Distributed Application Layer: Adaptive Mapping of Multiple Streaming Applications onto On-Chip Many-Core Systems

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Computer Engineering and Networks Laboratory

Lars Schor, Hoeseok Yang, Iuliana Bacivarov, Devendra Rai, and Lothar Thiele

Computer Engineering and Networks Laboratory, ETH Zurich, CH-8092 Zurich, Switzerland
firstname.lastname@tik.ee.ethz.ch

Problem

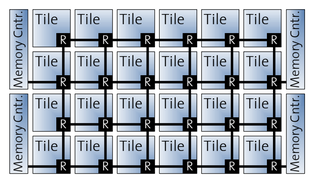
Software

- High performance requirements
- Multiple applications running in parallel
- Examples: smartphones, in-car entertainment systems

How to program and design such a system so that it is *analyzable* and *efficient*?

Hardware

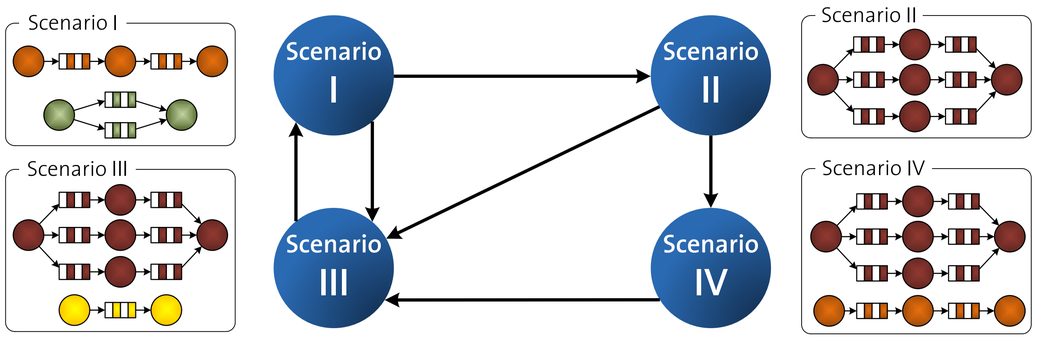
- Hundreds of cores
- Multiple communication layers
- Heterogeneous



Intel's SCC processor

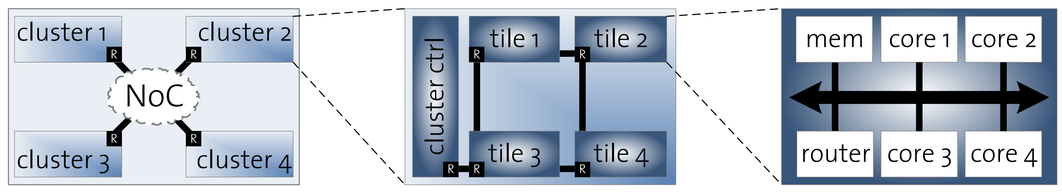
Specification

Workload Specification



- Execution scenarios to deal with multiple applications
- Each application is specified as a Kahn process network

Architecture Specification



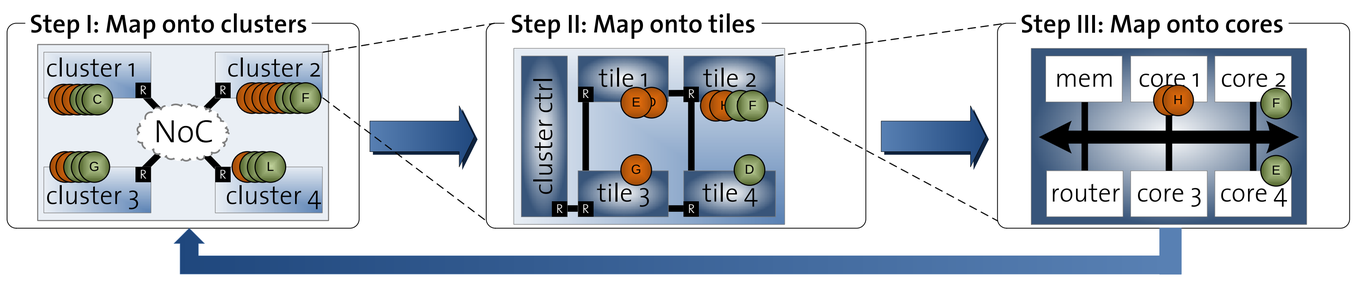
- Hierarchically organized
- Non-uniform memory access (NUMA) design
- Examples:
  - Intel SCC / Xeon Phi
  - STHorm (P2012)
  - NVIDIA Fermi architecture

Design-Time

- Mapping optimization:
  - *Objective*: minimize average power consumption
  - *Constraints*: real-time guarantees (including utilization, delay, ...)
- Basic idea:
  - Calculate an optimal mapping for each application and scenario
  - Hierarchical decomposition to improve the scalability

Scenario III / Scenario IV / Scenario II / Scenario I

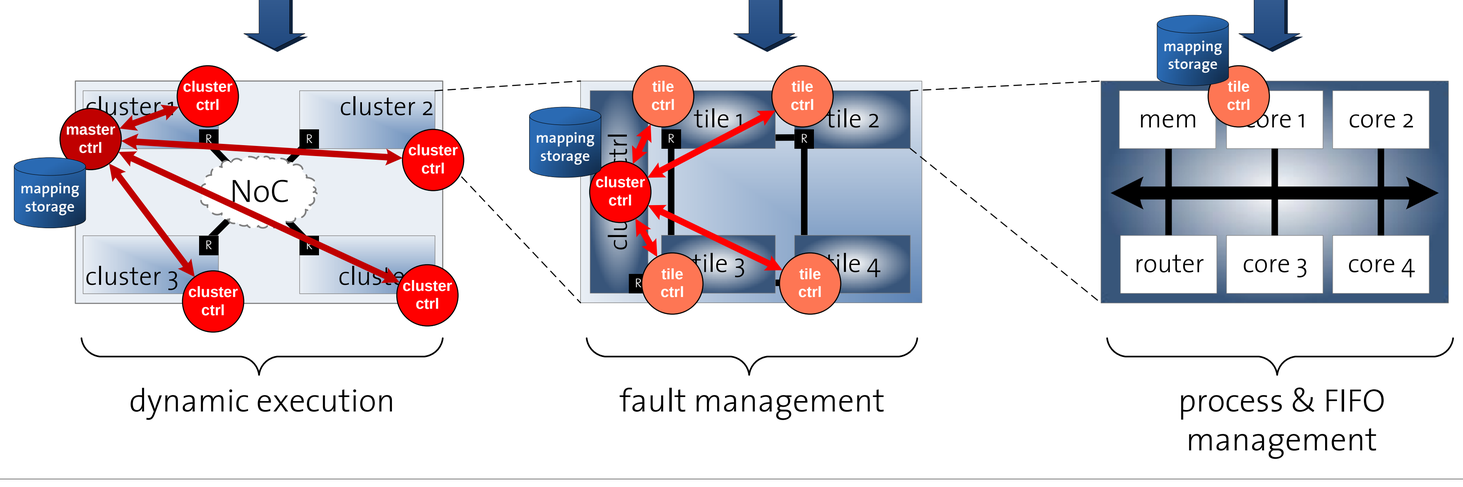
Hierarchical Decomposition:



- Assign processes to clusters
- To evaluate the objectives: characteristic of a cluster is speculated
- Individually analyze each cluster
- Again, characteristic of a tile is speculated
- Individually analyze each tile
- Feedback loop to improve the mapping

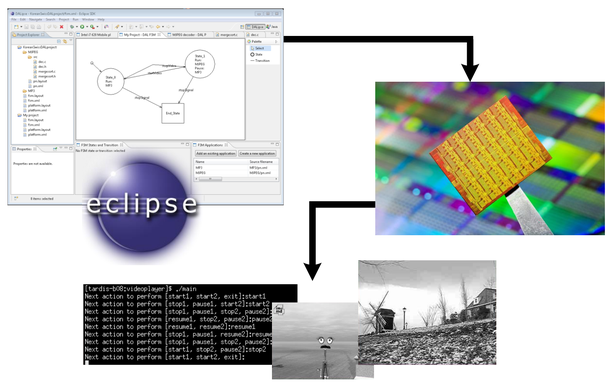
Run-Time

- Hierarchically organized run-time manager:
  - One controller per communication layer
  - Each controller has an individual database with its relevant mapping information
  - Events processed by the first controller that can handle the event
- Fault management:
  - Mapping towards virtual architecture
  - Redundant tiles to remap the processes



Evaluation

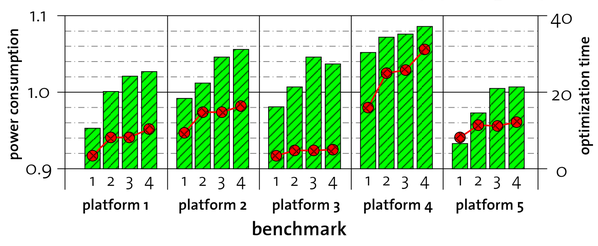
Deployment



Fully automated tool chain targeting Intel's SCC processor

Hierarchical Decomposition

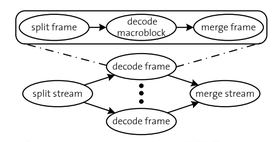
- How does hierarchical decomposition affect the performance of the mapping strategy?
- Performance of holistic mapping optimization relative to hierarchical mapping optimization:



Setup:
[1,16] scenarios
[40,160] processes
[16,64] cores

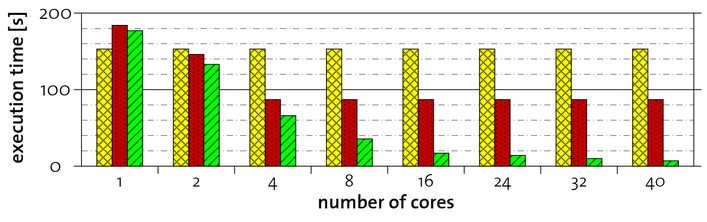
Different Levels of Parallelism

- Motion-JPEG (MJPEG) decoder application



Process network of the MJPEG decoder application.

- How does the degree of parallelism affect the throughput?



References:

- L. Schor, I. Bacivarov, D. Rai, H. Yang, S.-H. Kang, and L. Thiele: "Scenario-Based Design Flow for Mapping Streaming Applications onto On-Chip Many-Core Systems", Proc. Int'l Conf. on Compilers Architecture and Synthesis for Embedded Systems (CASES), Tampere, Finland, p. 71-80, Oct. 2012.
- S.-H. Kang, H. Yang, L. Schor, I. Bacivarov, S. Ha, and L. Thiele: "Multi-Objective Mapping Optimization via Problem Decomposition for Many-Core Systems", Proc. IEEE Symposium on Embedded Systems for Real-Time Multimedia (ESTIMedia), Tampere, Finland, p. 28-37, Oct. 2012.

Acknowledgement:

- EU FP7 project EURETILE (grant number 247846)
- Korean-Swiss Science and Technology Cooperation Program
- National Research Foundation of Korea (NRF-2011-357-D00213)
- Intel Doctoral Student Honor Programme

http://www.tik.ee.ethz.ch/~euretile