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

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

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

Workload Specification
- Scenario I
- Scenario II
- Scenario III
- Scenario IV
- 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

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

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

Deployment
- Fully automated tool chain targeting Intel’s SCC processor

Different Levels of Parallelism
- Motion-JPEG (MJPEG) decoder application
- How does the degree of parallelism affect the throughput?

References:

Acknowledgement
1. EU FP7 project CLARITY (grant number 248464)
2. Korean Science and Technology Cooperation Program
3. National Research Foundation of Korea (NRF-2012001-032 G003)
4. Intel Doctoral Student Honor Programme

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