Designing CMPs: So Many Options So Little Time

Polychronis Xekalakis

University of Edinburgh
CMPs: Why all the Fuss??

- We need more performance!
- Frequency cannot be scaled any more without burning the chip
- Big fat cores cannot extract more ILP
- Easiest way forward ... CMPs:
  - Replication is easy
  - Exploit Thread Level Parallelism
  - Increase throughput in multiprogramming environments
Design Time: A First Order Constraint

- Architects traditionally relied on simulators to understand the behaviour of the systems they build ... !
- Fully simulating is impossible
  - accurate simulation still very slow
  - number of simulations grows very quickly with number of parameters
  - slowdown almost linear with number of cores to simulate
- For Hetero systems, things are even worse
- Use Predictive Modeling!
Solving Hard Problems with Simple Tools

- Predictive modeling is essentially a simple supervised learning problem
- Train an ANN using a small number of simulations and predict the rest of the space
- Very simple solutions give very small errors ~ 2.9%
- But we can do even better, train on some of the programs and predict for the rest for a small hit in accuracy ~ 4.9%
Predictive Modeling Impact

Fraction of the space simulated and predicted shown for

(a) full simulation with no prediction
(b) per program prediction and
(c) cross program prediction.
Trading Accuracy with Speedup

- For Splash2 and a subset of TLS Spec Int.
  - 2.9% MAE for a speedup with an OM of 6
  - 4.9% MAE for a speedup with an OM of 8
Key Results - Observations

- Predictive modeling is a useful tool:
  - Not a hard-enough problem for ML people
  - But very useful for architects 😊
- Programs are not so different !!
  - Even explicitly parallel apps are similar with speculatively parallelized ones (TLS)
- Leveraging on this, one can gain an 8 orders of magnitude speedup by sacrificing 4.9% on accuracy
Current Work

- Currently working on designing heterogeneous systems
- The design problem gets bigger with the number of different types of cores we allow
- Mapping now plays a critical role and has to be considered as well
- Thus we try to design both the chip and the mapping heuristics to make use of it
Designing CMPs: So Many Options So Little Time

Polychronis Xekalakis
p.xekalakis@ed.ac.uk
University of Edinburgh