Optimized Algorithms for Network-on-Chip Application Mapping

Author: **Ciprian Radu**, PhD candidate
Prof.univ.dr.ing. Lucian Vinţan, m.c. ASTR, PhD supervisor

Computer Engineering Department

Lucian Blaga University of Sibiu, Romania

Outline

- Scope and objectives
- Network-on-Chip architectures
- Network-on-Chip application mapping
- UniMap framework
- Benchmarks
- Optimized Simulated Annealing
- Domain-knowledge Evolutionary Algorithms
- Application driven ADSE for SoCs
- Conclusions and further work

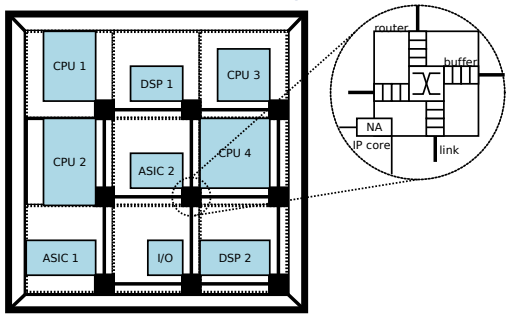
Scope and objectives

- Evaluation and optimization of Network-on-Chip application mapping algorithms
  - Develop a unified framework
  - Optimize Simulated Annealing
  - Optimize Evolutionary Algorithms
  - Application driven ADSE for SoCs

Outline

- Scope and objectives
- Network-on-Chip architectures
- Network-on-Chip application mapping
- UniMap framework
- Benchmarks
- Optimized Simulated Annealing
- Domain-knowledge Evolutionary Algorithms
- Application driven ADSE for SoCs
- Conclusions and further work

Network-on-Chip architectures

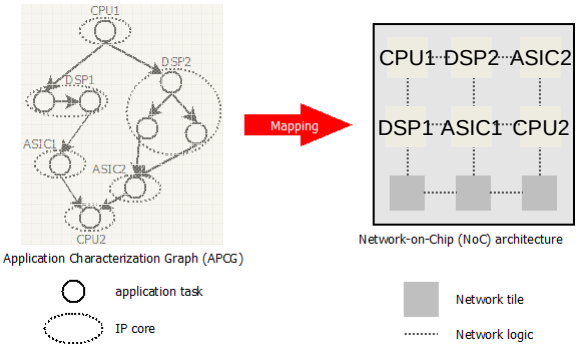


- Structured wiring
- Modularity
- Scalability
- Reliability (FT)
- Productivity
- Data abstraction
- Network modeling

Outline

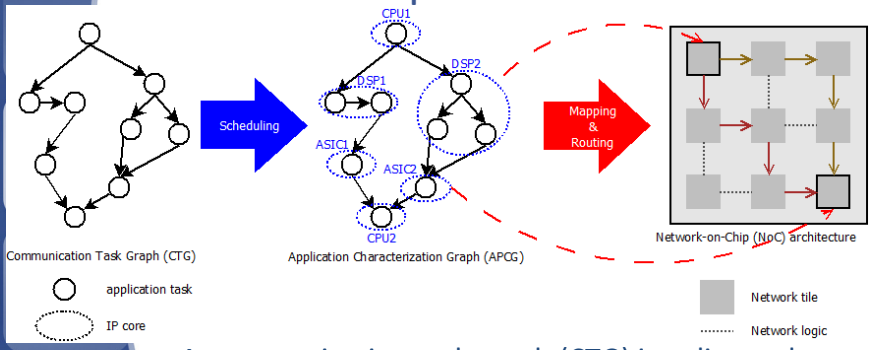
- Scope and objectives
- Network-on-Chip architectures
- Network-on-Chip application mapping
- UniMap framework
- Benchmarks
- Optimized Simulated Annealing
- Domain-knowledge Evolutionary Algorithms
- Application driven ADSE for SoCs
- Conclusions and further work

NoC application mapping



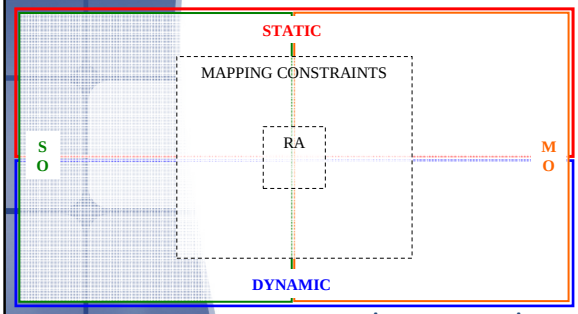
- An Application Characterization Graph (APCG) is a directed graph, G = G(C, A):
  - Each vertex $c_i \in C$ is an IP core
  - Each directed arc $a_{i,j} \in A$ marks the communication from c_i to c_j
- Application tasks (threads) are already assigned to IP cores

Application mapping, in relation to other NoC problems



- A communication task graph (CTG) is a directed graph G' = G'(T, A')
  - Each vertex $t_i \in T$ is an application task (thread)
  - Each directed arc $a_{i,j} \in A$ marks the communication from t_i to t_j

My taxonomy for application mapping algorithms



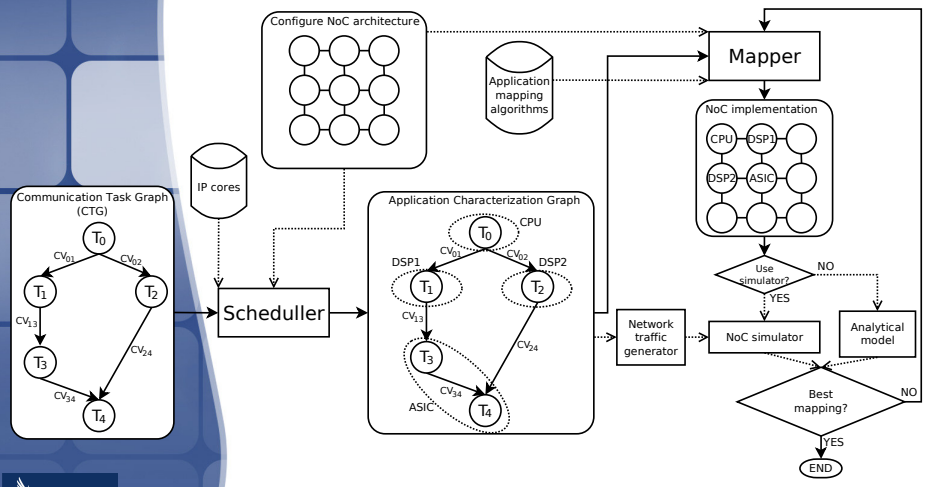
- mapping type
  - static
  - dynamic
- optimization goals
  - single objective (SO)
  - multiple objectives (MO)
- mapping constraints
  - with one or more mapping constraints
  - without any mapping constraint
- routing awareness (RA)
  - generates routes while mapping
  - does not generate routes

Outline

- Scope and objectives
- Network-on-Chip architectures
- Network-on-Chip application mapping
- UniMap framework
- Benchmarks
- Optimized Simulated Annealing
- Domain-knowledge Evolutionary Algorithms
- Application driven ADSE for SoCs
- Conclusions and further work

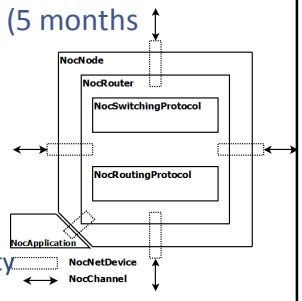
UniMap

https://code.google.com/p/unimap/

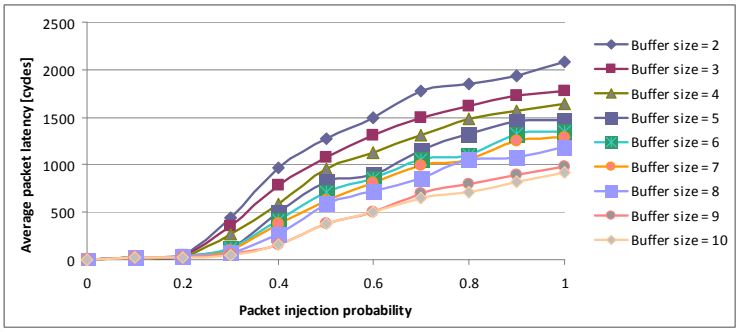


UniMap: the developed ns-3 NoC simulator

- Started at Augsburg University (5 months external research stage)
- Parameters
  - Traffic patterns
  - Network topology (k-ary d-cube)
  - Packet size
  - Packet injection rate & probability
  - Buffer size
  - Switching mechanism (SAF, VCT, Wormhole)
  - Routing protocol (DOR, SLB, SO)
- Estimates power consumption and area (using state of the art ORION 2.0)
- Network traffic generator for real applications



Experimental results with ns-3 NoC



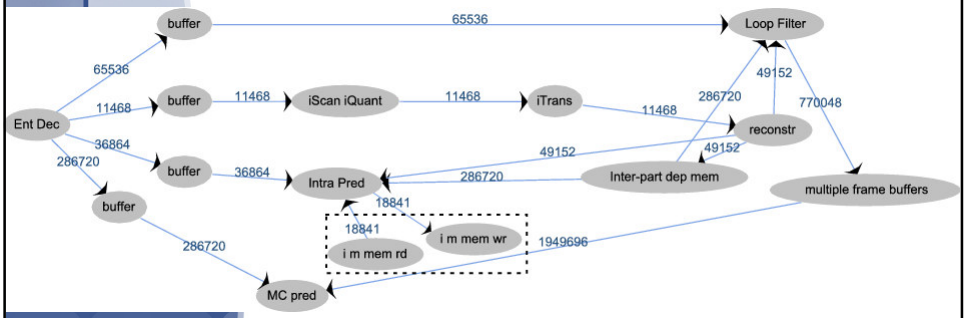
- 2x2x2x2 Hypercube NoC architecture
- DOR routing
- Wormhole switching
- Packet size = 9 flits
- Simulation for 10000 cycles (1000 warm-up cycles)

Outline

- Scope and objectives
- Network-on-Chip architectures
- Network-on-Chip application mapping
- UniMap framework
- Benchmarks
- Optimized Simulated Annealing
- Domain-knowledge Evolutionary Algorithms
- Application driven ADSE for SoCs
- Conclusions and further work

Benchmarks

- UniMap integrates the E3S benchmark suite + MMS, PIP, MPEG-4, MWD, VOPD, H.264



Outline

- Scope and objectives
- Network-on-Chip architectures
- Network-on-Chip application mapping
- UniMap framework
- Benchmarks
- Optimized Simulated Annealing
- Domain-knowledge Evolutionary Algorithms
- Application driven ADSE for SoCs
- Conclusions and further work

Optimized Simulated Annealing (OSA)

- Derived from a general energy- and performance-aware SA
- Employs and adapts some of the best SA practices for task scheduling
- Uses an application- and network-based exploration of the search space
- **Designed by me**
- The IP cores are implicitly and dynamically clustered using knowledge about communication demands

OSA characteristics

- Mapping cost: $E_{bit}^{p_i,p_j} = n_{hops} E_{R_{bit}} + (n_{hops} - 1) E_{L_{bit}}$
- Geometric annealing schedule: $T = T_0 \cdot q^{\left\lfloor \frac{i}{L} \right\rfloor}$, $q \in (0,1)$
- The *number of iterations per temperature level* (L) is application and network aware

$$L_{OSA} = C_n^2 - C_{n-c}^2 = \frac{c(2n-c-1)}{2} \quad , c, n \in N, n \geq c$$

$$L_{SA} = 100n^2$$

- OSA speedup over SA:

$$S = 1 - \frac{L_{OSA}}{L_{SA}} \overset{n=c}{\Rightarrow} \lim_{n \to \infty} S = 99.5\%$$

OSA characteristics

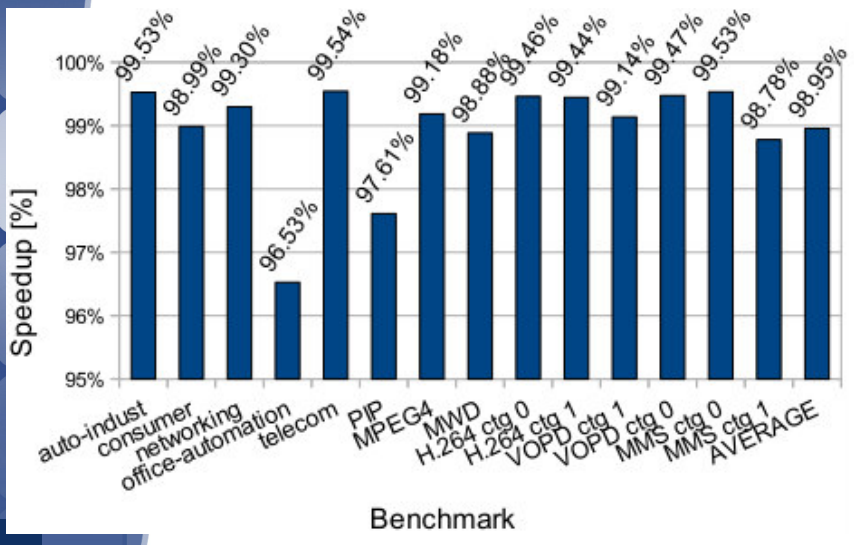
- Acceptance function: $P(\Delta C) = \frac{1}{1 + e^{\frac{\Delta C}{C_0 T}}}$
- PDF-based swapping
  - Replaces SA's (uniformly) random core swapping, with a core swapping based on Probability Density Functions (PDF)
  - Makes OSA to be application and NoC aware
  - Leads to dynamic and implicit core clustering

$$P[SelectedCore = i] = \frac{1}{c} + \frac{T}{T_0}\left(\frac{coreToComm_i}{totalToComm} - \frac{1}{c}\right)$$

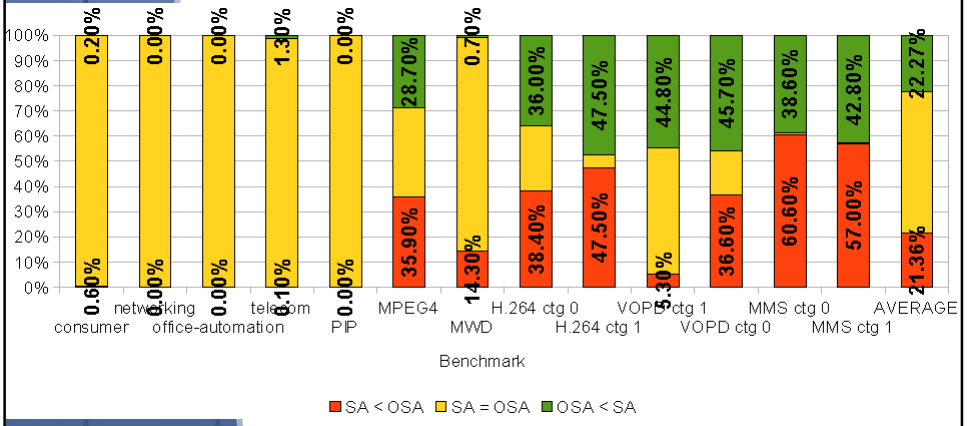
$$P[c_i \leftrightarrow c_j] = \frac{comm_{ij}}{totalComm}$$

- Stopping condition: coupled temperature and rejection threshold

OSA speedup over SA



OSA mappings cost, compared to SA mappings cost



Simulations on bigger 2D meshes

- 84 cores (10x9 NoC)
  - SA -> 70 hours; OSA -> 9.37 minutes; BB -> 6.33 minutes
  - OSA best mapping is 0.09% worse than SA's BB best mapping is 76% worse than OSA's
- 131 cores (12x11 NoC)
  - OSA -> 51 minutes (15% slower than BB)
  - BB best mapping is 79.4% worse than OSA's
- 215 cores (15x15 NoC)
  - OSA -> 8.4 hours (50% slower than BB)
  - BB did not find any solution!

Outline

- Scope and objectives
- Network-on-Chip architectures
- Network-on-Chip application mapping
- UniMap framework
- Benchmarks
- Optimized Simulated Annealing
- Domain-knowledge Evolutionary Algorithms
- Application driven ADSE for SoCs
- Conclusions and further work

Domain-knowledge Evolutionary Algorithms

- We developed an Elitist energy- and performance-aware Genetic Algorithm (EGA)
- We also used an Elitist Evolutionary Strategy (EES)
- Crossover operators
  - PB, PMX (standard for permutation problems)
  - NPB, MS (use domain-knowledge; **designed by us**)
- Mutation operators
  - Swap (standard for permutation problems)
  - OSA (→ metaheuristic)

NoC Position Based (NPB) crossover

- Context-aware Position Based crossover
- Keeps fixed the hot spot cores
- Parents are evaluated as

$$\text{cost} = \sum_{\substack{i,j \in C \\ i \neq j}} \text{vol}(c_i, c_j) \cdot \text{distance}(c_i, c_j)$$

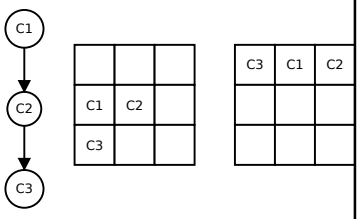
- The parent that better maps the hot spot cores is selected

Mapping Similarity (MS) crossover

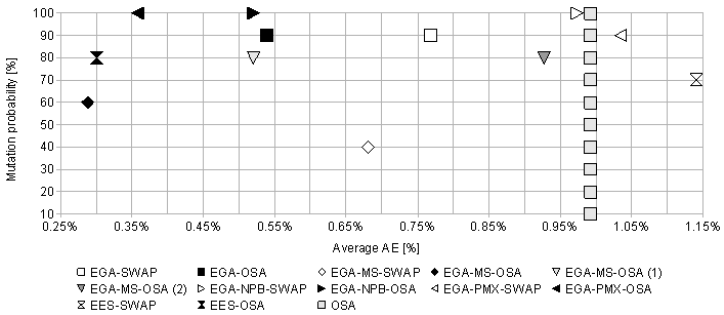
- Identifies topological similarities between two (parent) mappings and replicates them in the offspring
- Not similar cores are greedy mapped

$$D[i] = \sum_{\substack{i,k \in C \\ i \neq k}} d(i,k)$$

$$S[i] = \begin{cases} 1, & D_1[i] = D_2[i] \\ 0, & D_1[i] \neq D_2[i] \end{cases} \quad ,i \in C$$



Algorithms comparisons: best mapping cost

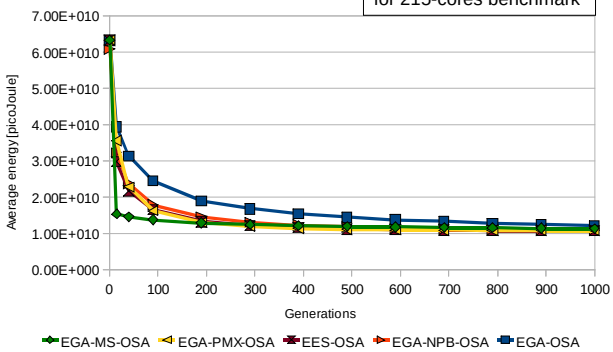


$$AE_{b_m} = \frac{x_{b_m} - \min\{X_b^{'}\}}{x_{b_m}}$$

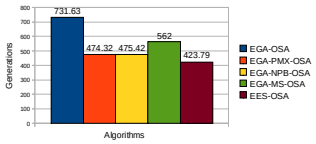
$m \in \{10\%, 20\%, ..., 100\%\}$ B = benchmark set

Algorithms comparisons: convergence

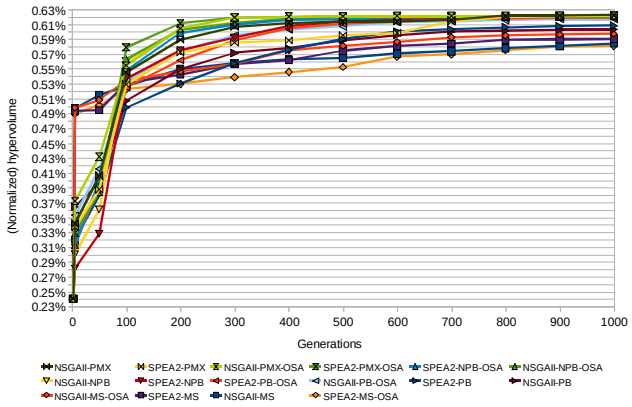
Algorithms' convergence for 215-cores benchmark



generations to reach best solution (average on all benchmarks)



Multi-objective evaluations

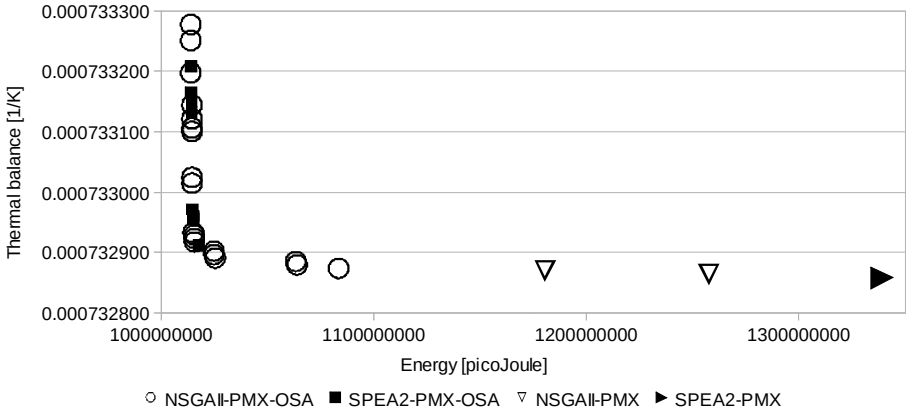


- 2 contradictory objectives
  - energy
  - thermal balance

$$\text{Thermal fitness} = \frac{1}{\max\{submatrices\ sums\}}$$

- UniMap generates a NoC floorplan for HotSpot

Multi-objective evaluations

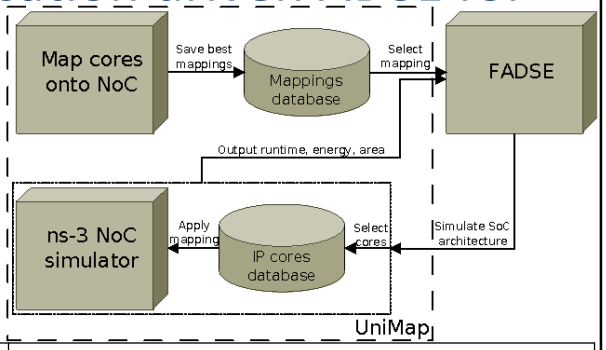


Outline

- Scope and objectives
- Network-on-Chip architectures
- Network-on-Chip application mapping
- UniMap framework
- Benchmarks
- Optimized Simulated Annealing
- Domain-knowledge Evolutionary Algorithms
- Application driven ADSE for SoCs
- Conclusions and further work

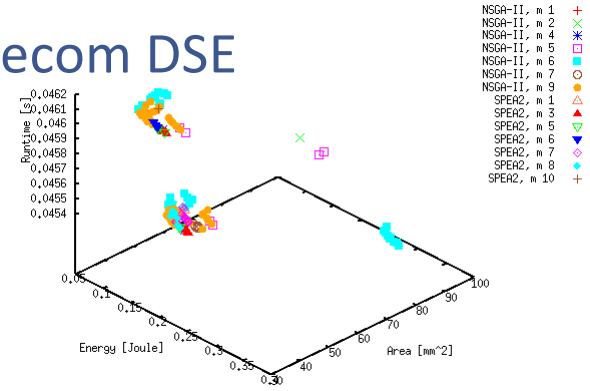
Application driven ADSE for SoCs

Benchmark	Search space size
telecom	$1.35 \cdot 10^{43}$
MPEG-4	$3 \cdot 10^{22}$
H.264 (CTG 0)	$4 \cdot 10^{28}$
VOPD (CTG 0)	$4 \cdot 10^{28}$



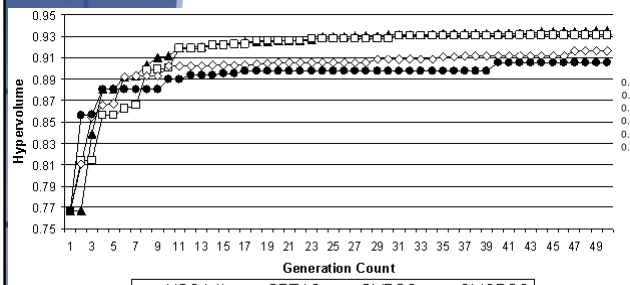
Benchmark		Core types	NoC parameters				
Name	Cores		Frequency [MHz]	Buffer size [flits]	Flit size [bytes]	Packet size [flits]	Routing
telecom	30	20	100, 200, ..., 1000	1, 2, ..., 10	4, 8, 16, ..., 256	2, 3, ..., 10	XY, YX
MPEG-4	12	34					
H.264 (CTG 0)	16	34					
VOPD (CTG 0)	16	34					

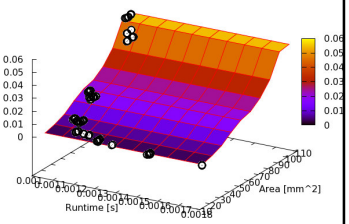
Telecom DSE

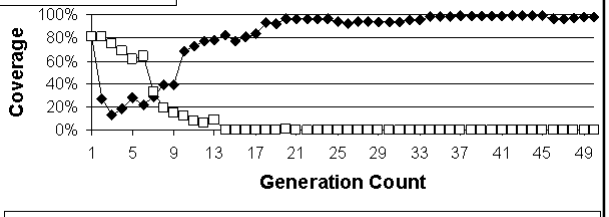


Objective	Algorithm	Mapping	NoC parameters					SoC energy [Joule]	SoC area [mm^2]	Application runtime [ms]
			Frequency [MHz]	Buffer size [flits]	Flit size [bytes]	Packet size [flits]	Routing			
Energy	NSGA-II	6	100	4	4	10	YX	0.095159	50.113	46.1144
Area	SPEA2	5	200	1	4	10	XY	0.158177	37.366	46.1132
Area	SPEA2	3	400	1	4	10	YX	0.167928	37.366	46.1111
Runtime	NSGA-II	6	900	4	32	6	YX	0.341914	81.227	45.4

MPEG-4 DSE







Outline

- Scope and objectives
- Network-on-Chip architectures
- Network-on-Chip application mapping
- UniMap framework
- Benchmarks
- Optimized Simulated Annealing
- Domain-knowledge Evolutionary Algorithms
- Application driven ADSE for SoCs
- Conclusions and further work

PhD main contributions

- UniMap
  - Real applications
  - Mapping algorithms library
  - Network-on-Chip simulator
  - Runs of HPC systems
- Optimized Simulated Annealing
  - 99% speedup from a general SA (without losing solution quality)
  - Feasible for NoCs larger than 10x10

PhD main contributions

- Single objective & multi-objective evaluations of Evolutionary Algorithms
  - MS improves convergence speed
  - An EA-OSA metaheuristic gives better results
- Application driven DSE workflow for SoC architectures
  - Best analytical mappings are not necessarily the best when using a simulator
  - GAs give better results than PSOs
  - PSOs converge faster than GAs

Further work

- Extract traffic patterns from shared memory and message passing parallel applications
- Research if graph morphism theory can applied to NoC application mapping
- Automatic Design Space Exploration for High Performance Computing Systems

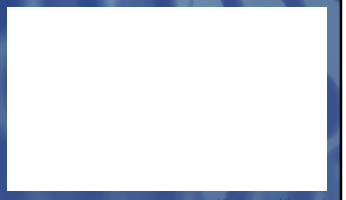
Papers

- **Ciprian Radu**, Lucian Vinţan, **UNIMAP: UNIFIED FRAMEWORK FOR NETWORK-ON-CHIP APPLICATION MAPPING RESEARCH**, Acta Universitatis Cibiniensis – Technical Series, "Lucian Blaga" University of Sibiu, Romania, ISSN 1583-7149, May 2011, Sibiu, Romania.
- **Ciprian Radu**, Lucian Vinţan, **Optimized Simulated Annealing for Network-on-Chip Application Mapping**, Proceedings of the 18th International Conference on Control Systems and Computer Science (CSCS-18), Politehnica Press, pp. 452-459, ISSN 2066-4451, 24 - 27 May 2011, Bucharest, Romania. **Selected for publishing in an Elsevier journal**.
- **Ciprian Radu**, Lucian Vinţan, **Towards a Unified Framework for the Evaluation and Optimization of NoC Application Mapping Algorithms**, Sixth International Summer School on Advanced Computer Architecture and Compilation for Embedded Systems (ACACES), Academic Press, Ghent, Belgium, pp. 163-166, ISBN 978-90-382-1631-7, July 14, 2010, Terrassa (Barcelona), Spain.
- **Ciprian Radu** and Lucian Vinţan, **Optimizing application mapping algorithms for NoCs through a unified framework**, In Proceedings of the 9-th IEEE RoEduNet International Conference, Sibiu, Romania, pp. 259 - 264, ISBN 978-1-4244-7335-9, 24-26 June 2010. IEEE Computer Society. Indexed IEEE, ISI, Scopus
- **Ciprian Radu**, Horia Calborean, Adrian Florea, Arpad Gellert, Lucian Vinţan, **Exploring some multicore research opportunities. A first attempt**, Fifth International Summer School on Advanced Computer Architecture and Compilation for Embedded Systems (ACACES), Academic Press, Ghent, Belgium, pp. 151-154, ISBN 978-90-382-1467-2, July 2009, Terrassa (Barcelona), Spain.
- Adrian Florea, **Ciprian Radu**, Horia Calborean, Adrian Crapciu, Arpad Gellert, Lucian Vinţan, **Understanding and Predicting Unbiased Branches in General-Purpose Applications**, Buletinul Institutului Politehnic Iasi, Tome LIII (LVII), fasc. 1-4, Section IV, Automation Control and Computer Science Section, Zentralblatt MATH indexed, pp. 97-112, ISSN 1220-2169, "Gh. Asachi" Technical University, 2007, Iaşi, Romania.
- Adrian Florea, **Ciprian Radu**, Horia Calborean, Adrian Crapciu, Arpad Gellert, Lucian Vinţan, **Designing an Advanced Simulator for Unbiased Branches' Prediction**, Proceedings of 9th International Symposium on Automatic Control and Computer Science, ISSN 1843-665X, November 2007, Iaşi, Romania.
- **Ciprian Radu**, Horia Calborean, Adrian Crapciu, Arpad Gellert, Adrian Florea, **An Interactive Graphical Trace-Driven Simulator for Teaching Branch Prediction in Computer Architecture**, The 6th EUROSIM Congress on Modelling and Simulation, (EUROSIM 2007), ISBN 978-3-901608-32-2, 9-13 September 2007, Ljubljana, Slovenia (special session: *Education in Simulation / Simulation in Education I*).

Thank you!

- Stay in touch with my research
  http://webspace.ulbsibiu.ro/ciprian.radu/

- Other questions
  ciprian.radu@ulbsibiu.ro



http://www.rubiks.com/