The Case for Custom Parallel Memories: an Application-Centric Analysis

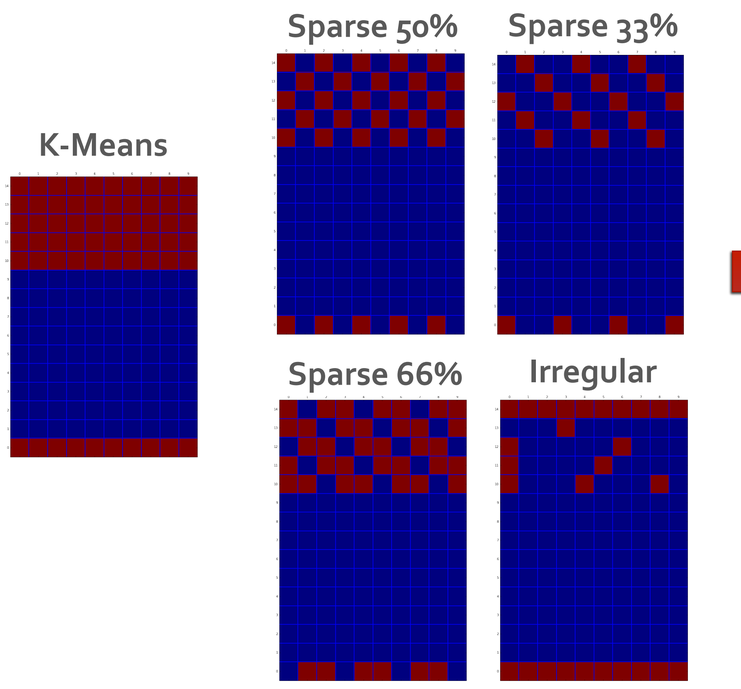
Giulio Stramondo, Cătălin Bogdan Ciobanu, Ana Lucia Varbanescu , Cees de Laat

University of Amsterdam

The main idea

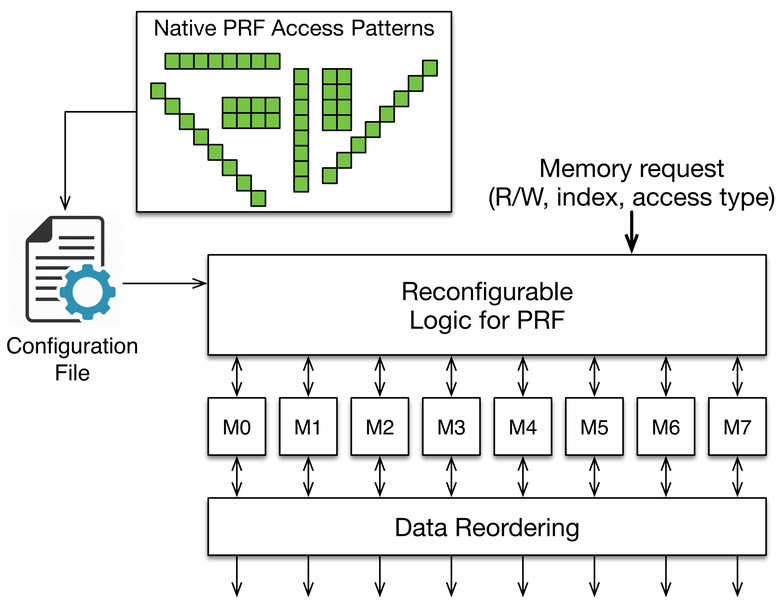
For many applications, the performance bottleneck is memory throughput. In this work, we propose to use FPGAs for building parallel custom memory systems to accelerate such applications.

Application Access Patterns



Parallel Reconfigurable Memory

We use the PRF [1] as a 2D reconfigurable scratchpad parallel memory.



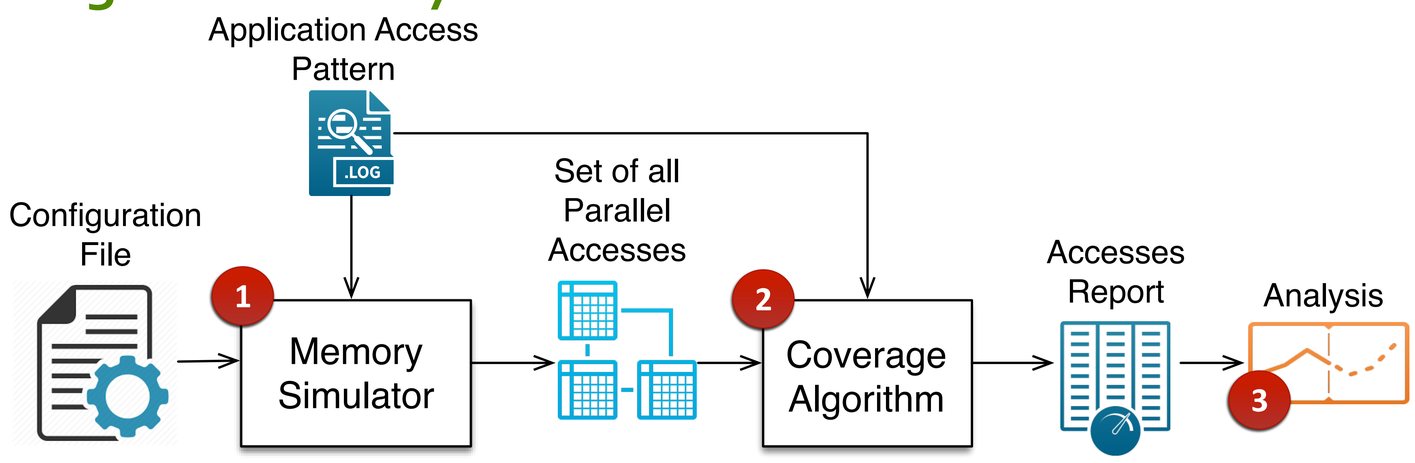
A configuration for the PRF specifies the number of memory used in parallel and (the subset of) the patterns to be supported.

Metrics

$$\text{Speedup} = \frac{\#\text{SequentialAccesses}}{\#\text{ParallelAccesses}}$$ Gain from using a parallel memory

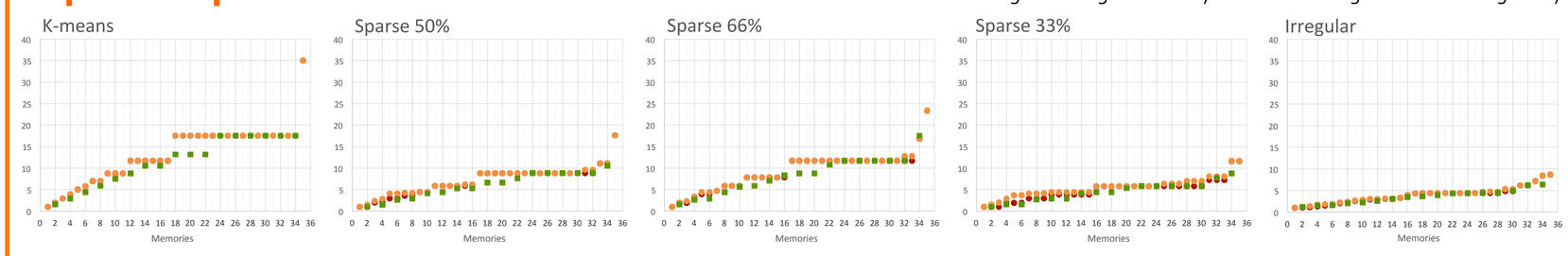
$$\text{Efficiency} = \frac{\text{ElementsReadByApplication}}{\text{ElementsReadByParallelMemory}}$$ Penalty from using a parallel memory

Configuration Analysis



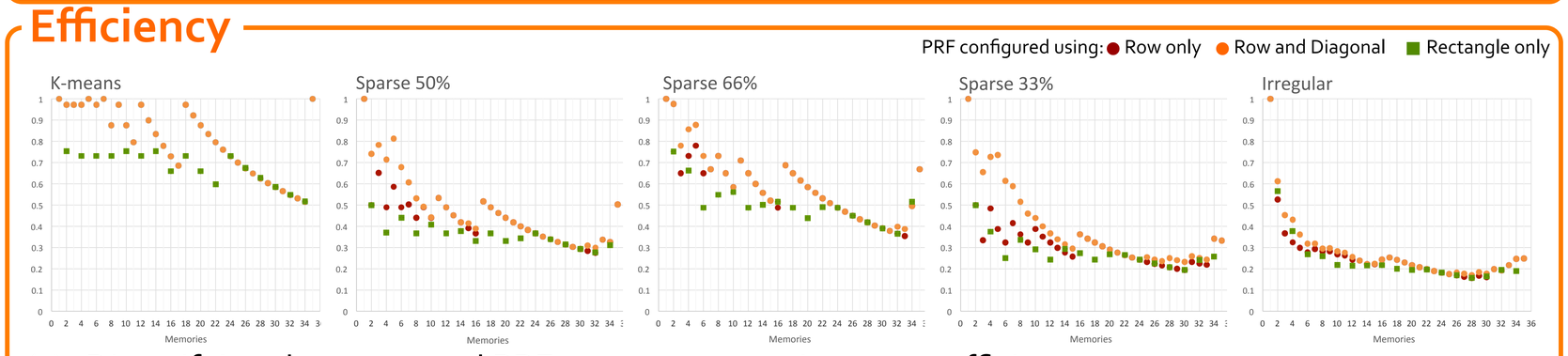
1. The memory simulator generates the possible parallel accesses. The number of elements accessible in each parallel access depends on the number of memories used in parallel.
2. The Coverage algorithm, which solves the minimum set cover using Integer Linear Programming (ILP), is used to find the minimum number of parallel accesses (as generated by the memory simulator) that cover the memory access pattern of the application.
3. The list of parallel accesses obtained by the ILP solver is used to compute **speedup** and **efficiency** of the analyzed configuration.

Speed-up



(1) Parallel memory systems provide speed-up in all cases.
(2) Combining multiple, different PRF access patterns is beneficial for speed-up.
(3) The "staircase" effect indicates that there are multiple configurations possible to achieve the same speed-up.

Efficiency



(1) Diversifying the supported PRF access patterns increases efficiency.
(2) The design of a parallel memory system should maximize efficiency to avoid useless memory ops; thus, for the same speed-up, the design with the highest efficiency should be selected.

Take Home Message

Using parallel memories is not trivial. Our approach offers an integrate semi-automatic approach to customize the memory system for your application, and evaluate its potential performance.

More info ?

Get in touch!
g.stramondo@uva.nl
c.b.ciobanu@uva.nl
a.l.varbanescu@uva.nl

EXTRA
www.extrahpc.eu

This project has received funding from the EU Horizon 2020 research and innovation programme under grant No 671653.The presentation has been partially supported by a SIGHPC grant.

References

[1] C. Ciobanu. **Customizable Register Files for Multidimensional SIMD Architectures.** PhD thesis, Delft University of Technology, Delft, Netherlands, March 2013.
[2] Krste Asanovic, Ras Bodik, Bryan Christopher Catanzaro, Joseph James Gebis, Parry Husbands, Kurt Keutzer, David A Patterson, William Lester Plishker, John Shalf, Samuel Webb Williams, et al. **The landscape of parallel computing research: A view from berkeley**. Technical report, Technical Report UCB/EECS-2006-183, EECS Department, University of California, Berkeley, 2006.
[3] Shuai Che, Michael Boyer, Jiayuan Meng, David Tarjan, Jeremy W Sheaffer, Sang-Ha Lee, and Kevin Skadron. Rodinia: **A benchmark suite for heterogeneous computing.** In Workload Characterization, 2009. IISWC 2009. IEEE International Symposium on, pages 44–54. IEEE, 2009.
[4] Richard M Karp. **Reducibility among combinatorial problems**. In Complexity of computer computations, pages 85–103. Springer, 1972.