

Proudly Operated by Battelle Since 1965

# Enabling the High Level Synthesis of Data Analytics Accelerators

MARCO MINUTOLI, VITO GIOVANNI CASTELLANA, ANTONINO TUMEO PACIFIC NORTHWEST NATIONAL LABORATORY RICHLAND, WA, USA

> MARCO LATTUADA ST MICROELECTRONICS LEGNANO, ITALY

> FABRIZIO FERRANDI POLITECNICO DI MILANO MILANO, ITALY

# New generation of *irregular* HPC applications

Pacific Northwest

Proudly Operated by Battelle Since 1965



**Big Science** 



#### **Complex Networks**



Language Understanding



#### **Bioinformatics**



#### Graph Databases



Pattern Recognition



#### **Community Detection**



**Knowledge Discovery** 

## **Massive Social Networks**

# facebook surpassed 1 billion active users



Pacific Northv

Operated by Battelle Since 1965

#### Statistics

- More than 1 billion active users, even more objects
- Average user has 130 friends and connected to 80 community pages, groups, and events
- Graph characteristics
  - Topology: Interaction graph is low-diameter and has no good separators
  - Irregularity: Communities are not uniform in size
  - Overlap: individuals are members of one or more communities

#### Sample queries:

- Allegiance switching: identify entities that switch communities.
- Community structure: identify the genesis and dissipation of communities
- Phase change: identify significant change in the network structure

# **Definition of Irregular Applications**



#### Irregularity in data structures

- Pointer- or linked-list based data structures such as graphs, unbalanced trees, unstructured grids
- Very poor spatial and temporal locality
  - Unpredictable data accesses
  - Fine grained data accesses
- Irregularity in control
  - Divergent branches
    - If (vertex==x) z; else k
- Irregularity in communication patterns
  - Unpredictable and fine grained communication
  - A consequence of irregularity in data structures and in control

# **Additional Characteristics**



#### Very large datasets

- Way more than what is currently available for single cluster nodes
- Very difficult to partition in a balanced way
- Large amounts of parallelism (e.g., each vertex, each edge in the graph)
- Usually, high synchronization intensity
  - Concurrent activities accessing the same elements of the data structures
- Datasets may be dynamically updated

# **Self-reinforcing Trend of FLOP-computing**



Proudly Operated by Battelle Since 1965

The HPC community builds systems for scientific simulations.



We need systems for data analysis, discovery, and inferencing.

# **Desirable System Features for Irregular Applications**



Proudly Operated by Battelle Since 1965

#### Multithreading

Tolerate, rather than reduce (with caches/locality) data access latencies

#### Global Address Space

No necessity to explicitly partition the dataset

#### Fine-grained synchronization

May need to lock a single memory word

Optimization: aggregation of fine-grained data accesses

# Exemplar Irregular & Data Analytics Application: Graph Databases

Pacific Northwest

- Promising solution to store large and heterogeneous datasets of these application fields
- Organize data in form of triples
  - Subject-predicate-object
  - Following the Resource Description Framework (RDF)
  - Set of triples represent a labeled, directed multigraph
- Queried through languages such as SPARQL
  - Fundamental operation is graph matching

# Graph Engine for Multithreaded Systems (GEMS)



- A software stack that implements a RDF (graph) database on a homogeneous commodity cluster
- Uses graph methods
- Converts SPARQL to graph pattern matching routines in C++
- Employs a custom Runtime (GMT – Global Memory and Threading) which provides:
  - Lightweight software multithreading
  - Message aggregation
  - Global address space

[V.G. Castellana, A. Morari, J. Weaver, A. Tumeo, D. Haglin, O. Villa, J. Feo:In-Memory Graph Databases for Web-Scale Data. IEEE Computer 48(3): 24-35 (2015)] Pacific Northwes

### **Query example**



#### Return the names of all persons owning at least two cars, of which at least one is a SUV



# **Source Code Example**



Proudly Operated by Battelle Since 1965

1 void 1 void 1 1 2 3 Edgy 4 for 5 dy 1 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1	<pre>mearch(Graph + graph, Nodeld var_2, Label p_var_3, Labelid _var_4, Labelid p_var_5, Labelid p_var_7, Labelid p_var_8, abelid p_var_9) { tin_degree_var_2 = getInEgreeigraph, var_2); tvar_2_1_inEdges(= getInEgreeigraph, var_2); tvar_2_1_inEdges(= uar_3); node; tvar_2_1_inEdges(i_var_3), node; tvar_2_1_inEdges(i_var_3), node; tvar_2_1_inEdges(i_var_3), node; tvar_2_1_inEdges(i_var_3), node; tvar_1_3; //el. with label *17* tur_1 = var_2_1_inEdges(i_var_7), node; tvar_1_1, inEdges(i_var_7), node; tvar_7 = var_1_1, inEdges(i_var_7), node; tvar_7 = var_1_1, inEdges(i_var_7), node; tvar_7 = var_1_1, inEdges(i_var_7), node; tf(var_7) = p_var_7) { size_t out_degree_var_6 = getOutEdges(graph, var_6); Edge + var_6_5, outEdges = getOutEdges(graph, var_6); Edge + var_6_5, outEdges = getOutEdges(graph, var_6); tf(var_7) = var_7) { tize_t out_degree_var_6 = getOutEdges(graph, var_6); tdur_9) //el. with label *ndfritype* var_9 = var_6_5, outEdges(i_var_9), node; tf(var_9 = var_6, outEdges(i_var_9), node; tf(var_9 = p_var_6) //el. with label *udr[l]Professor* var_6 = mortEdges(i_var_9), node; tf(var_9 = p_var_6) &amp; (var_6 = getOutEdges(graph, var_1); Edge + var_6_5 outEdges(i_var_9), node; tf((var_9 = p_var_6), 4(var_6 = p_var_6)) ( size_t out_degree_var_1) = getOutEdges(graph, var_1); Edge + var_1_7_outEdges(i_var_5), property; Nodeld var_5; //el. with label *udr:ype* var_5 = var_1.7_outEdges(i_var_5), property; Nodeld var_5; //el. with label *udr:ype* var_5 = var_1.7_outEdges(i_var_5), node; tize_t out_1.7_outEdges(i_var_5), node; tize_t out_1.7_outEdges(i_var_5), node; tize_t our_1.7_outEdges(i_var_5), node; tize_t our_1.7_outEdges(i_var_6), property; Nodeld var_5 = p_var_5) &amp; ki (var_4 = p_var_4))</pre>	<pre>void kernel(size_t i_var3, Edge v var_2,1_inSdges, Graph</pre>	1 23455789 101121314151617 189201222 23 24 25 26 27 28 29 30 313
		(b)	

(a)

Fig. 5: Pseudo code for the pattern matching routines of example query Q6

# Can we use FPGAs to accelerate Data Analytics and Irregular Applications?



- FPGAs provide an opportunity to implement application specific accelerators fast, and power efficient
- Flexibility challenge: implementing accelerators in RTL (Register Transfer Level) languages is complex and time consuming
  - High-Level Synthesis (HLS): synthesizing RTL from descriptions in higher level languages (e.g., C)
  - However, conventional HLS tools typically target Digital Signal Processing algorithms, i.e., typical "regular" applications
- ► We can accelerate GEMS by:
  - Implementing some parts of the runtime on FPGA
  - Directly synthesizing queries (i.e., graph pattern matching routines)
    - Synthesis time is not a limitation: queries do not change often, datasets do

## **GEMS on FPGAs**





- C code generated by GEMS SPARQL-to-C++ translator
- Code is then processed by Bambu, a High Level Synthesis tool from Politecnico di Milano
  - Heavily modified to support our new architectural templates
- Note: accelerating graph walks is a more complex problem than accelerating table operations

[V. G. Castellana, M. Minutoli, A. Morari, A. Tumeo, M. Lattuada, F. Ferrandi: High Level Synthesis of RDF Queries for Graph Analytics. ICCAD 2015]

[M. Minutoli, V. G. Castellana, A. Tumeo: High-Level Synthesis of SPARQL queries. SC15 poster]

# Challenges for HLS of Irregular Applications



Challenge 1: Coarse grained parallelism exploitation

- Most HLS approaches focus on exploiting ILP and do not support TLP specifications (expressed through parallel programming APIs such as OpenMP, CUDA, OpenCL)
- Concurrency and synchronization management
- Target architecture design: HLS flows usually generate Finite State Machines with Datapaths, which are inherently serial
- Challenge 2: Support for complex memory subsystems
  - Dynamic resolution of memory addresses
  - Pointer-based data structures
  - Memory consistency and synchronization
  - Barriers, Atomic memory operations
  - Distributed/multi-ported memories

# Solving Challenge 1: Parallel Distributed Controller



- From serial FSMD to a parallel distributed controller
- Architecture:
  - Set of communicating control elements, called Execution Managers (EMs)
    - Each EM establishes when an operation/task can start at runtime
    - Dynamic execution paradigm
    - Dedicated hardware (Resource Managers RMs) for checking:
      - Satisfaction of dependence constraints
      - Resource availability
    - Natural support for variable latency operations/tasks
      - ASAP execution

## **Example of Parallel Distributed Controller**







[V.G. Castellana, A. Tumeo, F. Ferrandi: An adaptive Memory Interface Controller for improving bandwidth utilization of hybrid and reconfigurable systems. DATE 2014.]

Pacific Northy

- Allows concurrent memory accesses on distributed/multi-ported shared memories
- Dynamically resolves memory addresses
- Manages concurrency and synchronization (supporting atomic operations)





#### Manages concurrency

- Provides memory scrambling (solves hotspot)
- Collects memory requests, and if their target addresses collide, it forwards them one at a time
- Each memory port is managed by a dedicated arbiter
- No delay penalties

#### Manages synchronization

- Directly implements atomic operations (e.g. atomic increment, compare and swap)
- Inside the interface, through dedicated hardware
- While running, atomic operations *lock* the associated memory port

# **Hierarchical Approach to HLS**





- A function is a module, and can contain other function modules
- Function modules can use parallel controller (e.g., launching iterations of a loop, where each iteration has a corresponding hardware kernel), or a FSMD module (e.g., the loop iteration itself)
- Allows generating accelerators for irregular codes with nested loops
  - Graph algorithms & queries

# Load Unbalancing in Queries

Iteration





# **Dynamic Task Scheduler**



Proudly Operated by Baffelle Since 1965



[Marco Minutoli, Vito Giovanni Castellana, Antonino Tumeo, Marco Lattuada, Fabrizio Ferrandi: Efficient Synthesis of Graph Methods: A Dynamically Scheduled. ICCAD 2016]

- The Task Queue stores tasks ready for execution
- The Status Register keeps track of resource availability
- The Task Dispatcher issues the tasks
- The Termination Logic checks that all tasks have been used

# **Experimental Evaluation**



	Single Acc.	Parallel	Dynamic	Speedup				
	# Cycles	Controller # Cycles	Scheduler # Cycles	Single Acc.	Parallel Controller			
Q1	5,339,286	5,176,116	5,129,902	1.04	1.01			
Q2	141,022	54,281	50,997	2.77	1.06			
Q3	5,824,354	1,862,683	1,805,731	3.23	1.03			
Q4	63,825	42,851	19,928	3.20	2.15			
Q5	33,322	13,442	9,016	3.70	1.49			
Q6	674,951	340,634	197,894	3.41	1.72			
07	1,700,170	694,225	492,280	3.45	1.41			

#### LUBM-1 (100k triples)

#### LUBM-40 (5M triples)

	Cinals Ass	Parallel	Dynamic	Speedup				
	# Cycles	Controller # Cycles	Scheduler # Cycles	Single Acc.	Parallel Controller			
Q1	1,082,526,974	1,001,581,548	287,527,463	3.76	3.48			
Q2	7,359,732	2,801,694	2,672,295	2.75	1.05			
Q3	308,586,247	98,163,298	95,154,310	3.24	1.03			
Q4	63,825	42,279	19,890	3.21	2.13			
Q5	33,322	13,400	8,992	3.71	1.49			
Q6	682,949	629,671	199,749	3,42	3.15			
Q7	85,341,784	35,511,299	24,430,557	3.49	1.45			

- Architectures integrating dynamic scheduling vs. solutions only integrating PC+MIC
- Dynamic Scheduling always provides higher performance
- In the majority of cases, speed ups are over 3 (with 4 accelerators)
- The design is also more area efficient: higher speed up than area overhead (also w.r.t. parallel controller)

## **Latest Extension: Temporal Multithreading**



Proudly Operated by Battelle Since 1965

- With Irregular Applications, we prefer to tolerate latency
  - Latency reduction through caches and localization ineffective
- Objective is to saturate the memory subsystem with parallel memory operations
  - Temporal, rather than spatial, multithreading provides interesting area/performance/utilization tradeoffs
- We extend the DTS design to perform context switching after memory operations
  - Memory interface further decoupled
  - Number of contexts is configurable



#### Svelto Methodology

## **Architectural Template Mapping**



Proudly Operated by **Battelle** Since 1965



		void alomic.update() (
	2	update_results();
	3	Friday and the second second
	4	void loop_iteration(size_t i,) {
	5	{} // loop body X
	6	atomic_update();
void top.function() {	7	{} // loop body Y
{} // code block A	8	) 10 10 10 10 10 10 10 10 10 10 10 10 10
#pragma omp parallel for	9	void parallel.loop() {
for (size t i = 0; i < N; ++i) {	10	for (size_t i = 0; i < N; ++i)
{} // loop body X	11	loop_iteration(i,);
#pragma omp atomic	12	1
update.results();	13	void top.function() {
() // loop body Y	14	() // code block A
}	15	parallel.loop();
{} // code block B	16	<pre>{} // code block B</pre>
1 contraction of the second second	17	)

(a) Example of OpenMP application to be synthesized.

2345

6

8

10

11

(b) Example of OpenMP application to be synthesized after HLS transformations.

You full to be a first start of a first of the first of the

# Multithreaded Architecture Template: worker and memory controller



Memory

 Architecture of a Single multithreaded worker

#### Architecture of the Top Memory Controller





Arbiter

# **First example – Sequential Accelerator**



- Create a sequential accelerator implementing tq4 specifying input/output testbench with xml file
- Hints:
  - Start from the bambu.sh script
  - Set the top function equal to search
  - Specify that the memories are outside the accelerator
  - Set the memory delay to an high value
  - Specify as testbench argument the xml that is in the directory

# **Solution**



bambu --compiler=I386\_GCC49 --std=c99 --experimentalset=BAMBU -03 -v3 -fno-delete-null-pointer-checks -channels-type=MEM\_ACC\_11 --memory-allocationpolicy=NO\_BRAM --device-name=xc7vx690t-3ffg1930-VVD -clock-period=10 -DMAX\_VERTEX\_NUMBER=26455 -DMAX\_EDGE\_NUMBER=100573 -DN\_THREADS=1 \${root\_dir}/common/atominIncrement.c \${root\_dir}/common/data.c -I\${root\_dir}/common/ \${root\_dir}/trinityq4/lubm\_trinityq4.c \ --top-fname=search \

- --mem-delay-read=20 --mem-delay-write=20 \
- --memory-allocation-policy=NO\_BRAM \
- --simulate --generate-tb=\${root\_dir}/test-1.xml

## **Second example – Parallel Accelerator**



- Create a paralell accelerator implementing tq4 specifying input/output testbench with xml file
- Configuration to be considered
  - 2 Copies of the kernel
  - 4 External memory banks
  - 2 Internal channels
  - No context switch (i.e., context-switch==1)

# **Solution**



bambu --compiler=I386 GCC49 --std=c99 --experimentalset=BAMBU -03 -v3 -fno-delete-null-pointer-checks -channels-type=MEM ACC 11 -- memory-allocation-policy =NO BRAM --device-name=xc7vx690t-3ffg1930-VVD --clockperiod=10 -DMAX VERTEX NUMBER=26455 -DMAX EDGE NUMBER=100573 -DN THREADS=2 \${root dir}/common/atominIncrement.c \${root dir}/common/data.c -I\${root dir}/common/ \${root dir}/trinityq4/lubm trinityq4.c --topfname=search --mem-delay-read=20 --mem-delay-write=20 --memory-allocation-policy=NO BRAM --simulate -generate-tb=\${root dir}/test-1.xml \ -num-threads=2 \ --memory-banks-number=4 \ --channels-number=2  $\setminus$ --context switch=1

# **Third example – Context Switch**



- Create a parallel accelerator implementing tq4 specifying input/output testbench with xml file
- Configuration to be considered
  - 2 Copies of the kernel
  - 4 External memory banks
  - 2 Internal channels
  - 4 Logic Threads x kernel

# **Solution**



bambu --compiler=I386 GCC49 --std=c99 --experimentalset=BAMBU -03 -v3 -fno-delete-null-pointer-checks -channels-type=MEM ACC 11 -- memory-allocation-policy =NO BRAM --device-name=xc7vx690t-3ffg1930-VVD --clockperiod=10 -DMAX VERTEX NUMBER=26455 -DMAX EDGE NUMBER=100573 -DN THREADS=2 \${root dir}/common/atominIncrement.c \${root dir}/common/data.c -I\${root dir}/common/ \${root dir}/trinityq4/lubm trinityq4.c --topfname=search --mem-delay-read=20 --mem-delay-write=20 --memory-allocation-policy=NO BRAM --simulate -generate-tb=\${root dir}/test-1.xml \ -num-threads=2 \ --memory-banks-number=4 \ --channels-number=2  $\setminus$ --context switch=4

## **Fourth example – Context Switch DSE**



- Create parallel accelerators implementing tq4 specifying input/output testbench with xml file using context switch
- Evaluate different configurations
  - Double the number of contexts
  - Double the number of memory banks
  - Etc.





- --memory-banks-number=4 --channels-number=2 --context\_switch=8
- --memory-banks-number=8 --channels-number=2 --context\_switch=8
- --memory-banks-number=16 --channels-number=2 --context switch=8

# **Design Space Exploration**

(a) Q1



(d) Q4

Proudly Operated by **Battelle** Since 1965



(c) Q3

(b) Q2







Query 593 with 4CH







### **Results comparison**



Parallel Dynamic Svelto Speedup Controller Scheduler PC DS # Cycles # Cycles # Cycles 1,001,581,548 287,527,463 269,158,569 3.72 1.07 Q1 Q2 2,801,694 2,672,295 2,422,525 1.16 1.10 Q3 98,163,298 95,154,310 81,911,448 1.20 1.16 42,279 Q4 19,890 18,128 2.33 1.10 Q5 13,400 8,992 8,555 1.57 1.05 Q6 629,671 199,749 171,689 3.67 1.16 Q7 35,511,299 24,430,557 21,509,718 1.65 1.14

	Parallel Controller		Dynamic Scheduler		Svelto		Svelto vs PC			Svelto vs DS					
	Freq. (Mhz)	LUTs (#)	Slices (#)	Freq. (Mhz)	LUTs (#)	Slices (#)	Freq. (Mhz)	LUTs (#)	Slices (#)	Freq. (%)	LUTs (%)	Slices (%)	Freq. (%)	LUTs (%)	Slices (%)
Q1	113.37	13,469	4,317	113.60	10,844	3,503	123.35	7,434	2,314	8.80	44.81	46.40	8.58	31.45	33.95
Q2	130.11	5,280	1,607	132.87	4,636	1,335	121.18	4,612	1,487	-6.86	12.65	7.47	-8.80	0.52	-11.39
Q3	114.53	13,449	4,308	116.92	10,664	3,467	110.56	7,378	2,390	-3.47	45.14	44.52	-5.44	30.81	31.06
Q4	122.97	7,806	2,399	118.68	6,175	1,918	133.14	5,712	1,765	8.27	26.83	26.43	12.18	7.50	7.98
Q5	138.31	5,750	1,738	114.51	5,330	1,578	153.28	4,776	1,524	10.82	16.94	12.31	33.86	10.39	3.42
Q6	113.26	10,600	3,426	118.68	8,125	2,633	117.90	6,112	1,983	4.10	42.34	42.12	-0.66	24.78	24.69
Q7	106.71	15,002	4,953	113.23	11,344	3,747	115.83	7,589	2,469	8.55	49.41	50.15	2.30	33.10	34.11

## Conclusions



- Identified challenges due to irregular behaviors in Data Analytics applications
- Introduced GEMS, our graph database for homogeneous clusters, which supports RDF and SPARQL
- Highlighted possible ways to accelerate SPARQL queries with custom accelerators in our framework
- Identified current limitations of High-Level Synthesis (HLS) for Data Analytics applications
- Presented architectural templates and methodologies for the HLS of Data Analytics applications
- Presented results of the synthesis of SPARQL queries with our methodology

# Thank you for your attention!



Proudly Operated by Baffelle Since 1965

#### Questions?

- antonino.tumeo@pnnl.gov
- vitoGiovanni.castellana@pnnl.gov
- marco.minutoli@pnnl.gov

fabrizio.ferrandi@polimi.it