Centre of Excellence in Information and Communication Technologies

High Performance FPGA Based BLAS accelerator

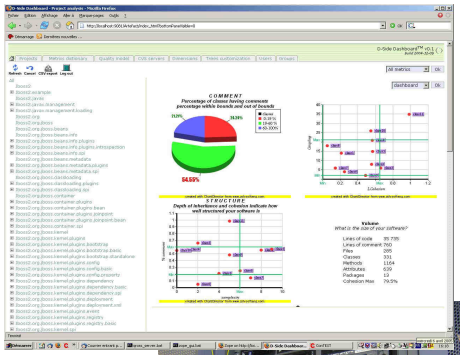
Lotfi Guedria

CETIC

- Research Center in Information and Communication Technologies
  - Created in 2001, by 3 Belgian Universities
  - Non-profit center, 27 researchers, 2.5 M€ budget
- Mission
  - Perform Applied Research & Technology transfer
  - Connecting Research Labs To Enterprises
  - Serve the Industry
    - Contribution to *Regional Economic Development*
- Initiated by: UCL Université catholique de Louvain; Facultés Universitaires Notre-Dame de la Paix; POLY TECH. MONS
- Supported by:

CETIC Research Areas

- **Software Engineering**
  - Software Process & Products Quality
    - Metrics, Monitoring, Impact on software processes
  - Requirements engineering
    - Security
    - Critical Systems Modelling
- **Distributed Technologies**
  - Grid Technologies, Cluster
    - Distributed data management
    - Collaborative technologies
  - Web data mining
    - Search Engine
    - Reverse Engineering
- **Embedded Systems**
  - Wireless Technologies
  - Co-design - FPGA



Sun microsystems Center of Excellence

Outline

- Introduction
- Theoretical Study
- Prototype Implementation
- Measured Results
- Limitation Analysis
- Conclusion

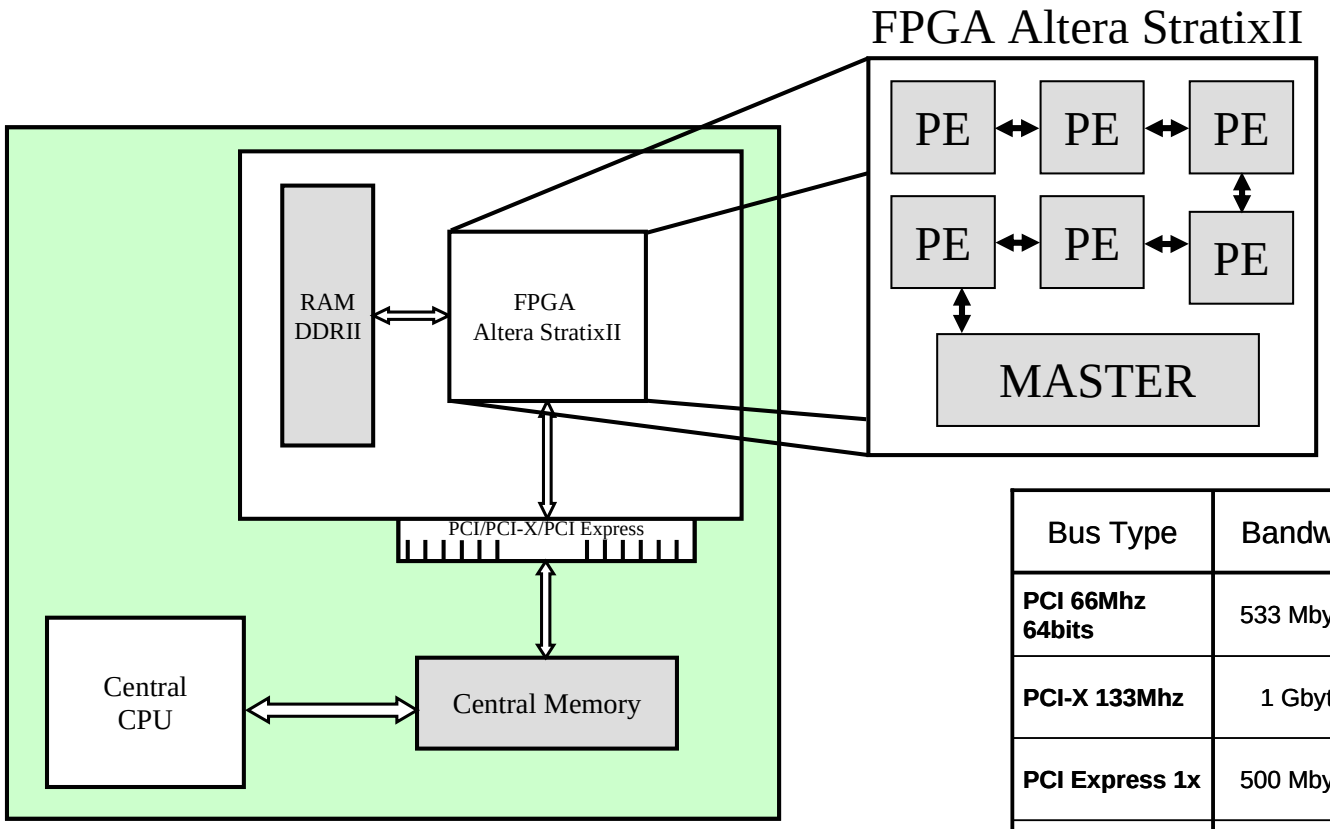
Project Description

- Study/development of a Hardware Accelerator for the BLAS library

- Specifications:
  - Installable in a common computer
  - Easy to use and to install
  - Transparent for the final user

→ Idea: Use a FPGA-based system to accelerate the CPU time spent executing BLAS operations

System Diagram



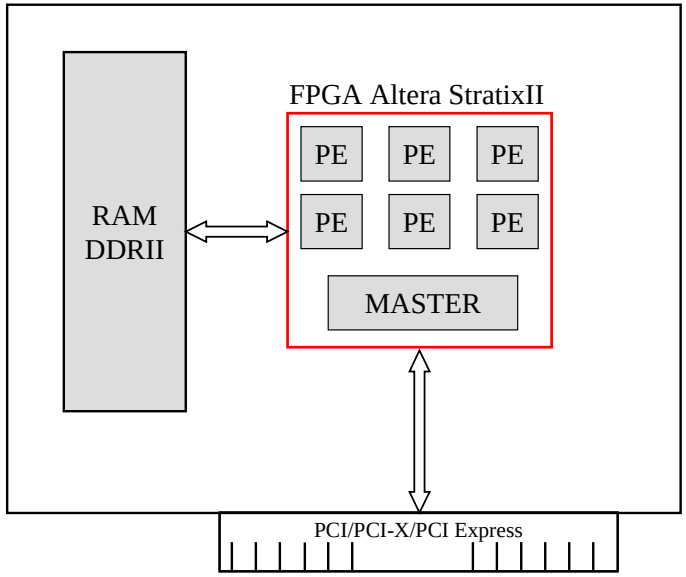
Host computer + FPGA board

Bus Type	Bandwidth
PCI 66Mhz 64bits	533 Mbytes/s
PCI-X 133Mhz	1 Gbyte/s
PCI Express 1x	500 Mbytes/s
PCI Express 4x	2 Gbytes/s
PCI Express 8x	4 Gbytes/s

Outline

- Introduction
- Theoretical Study
- Prototype Implementation
- Measured Results
- Limitation Analysis
- Conclusion

Performance Evaluation



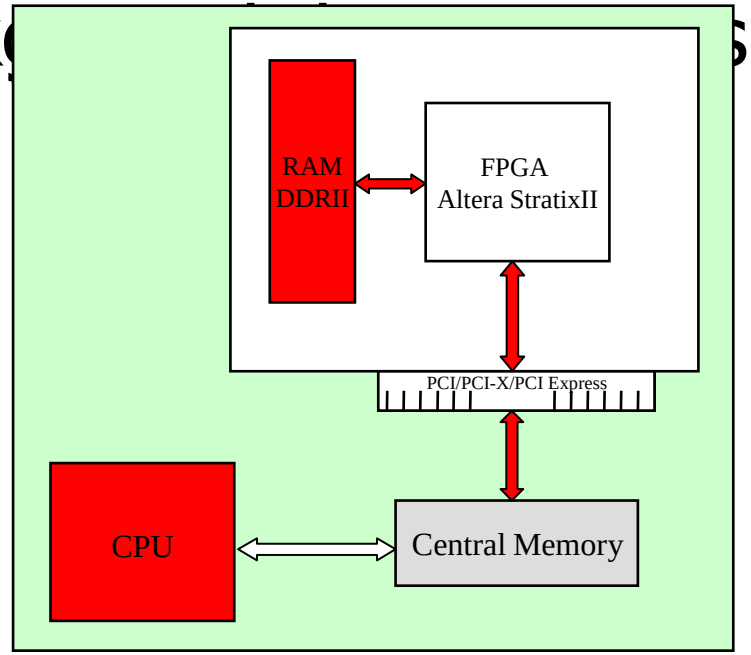
- FPGA Theoretical Raw Computation performance:

  Op x PE x Fmax [Gflops/s]

- Doesn't take into account the data transfers

Limitations

- FPGA performance (less likely to be the limiting factor)
- data transfers
- CPU back[illegible]sing



BLAS Operation Selection (1)

Address the Memory Accesses Limitations

1. At Hw level : Highest memory bandwidth possible
2. At algorithmic level : Maximise the number of operations performed per memory access

→ Va factor: [Ralf Gruber]

$$Va = O / Ma \quad [op/memory\ access]$$

O = Total number of operations

Ma = Total number of memory accesses

=> Operations maximising the Va factor will maximise the global performances

BLAS Operation Selection (2)

→ Implementation of the BLAS Level 3 matrix multiplication operation : « dgemm »

$O = 2 \times n^3$, $Ma = 2 \times n^2$ → $Va = n$

$$C_{ij} = \sum_{k=0}^{N} A_{ik} \cdot B_{kj}$$

with i = 1,…,M et j = 1,…,K

Outline

- Introduction
- Theoretical Study
- Prototype Implementation
- Measured Results
- Limitation Analysis
- Conclusion

FPGA Board Selection

- Board specifications :
  - ➢ One or more high density FPGA (≥60k LE)
  - ➢ High bandwidth communication interface (PCI Express 8x)

→ Selected FPGA board:

Gidel ProcStarII:

- StratixII 60k LE
- PCI 64 bits 66 Mhz

MAC Unit Implementation Evaluation

FPGA Altera StratixII 60

	ALTERA MAC IP EP2S60-3	OUR MAC IP EP2S60-3
ALUT	2451 / 48352	1539 / 48352
MULT 18x18	9 / 144	9 / 144
Fmax	143,74 Mhz	235,32 Mhz
Pipeline Stages	14	14

→144/9 = 16 PE in a StratixII 130

→2(FPGA) x 32 Flops x 200 Mhz = 12,8 Gflops/s

Matrix Multiplication Implementation

- Implementation of the complete design composed of 16 PE
- Limitations:
  - Design complexity:
    → Maximum achievable (placement): 14 PE in a StratixII 60
  - Memory controller performance:
    → FPGA design will run at 140 Mhz max

→ Matrix Multiplication performance with 14 PE:

2(FPGA) x 2 x 14 x 140 Mhz = 7,8 Gflops/s

Dgemm Implementation

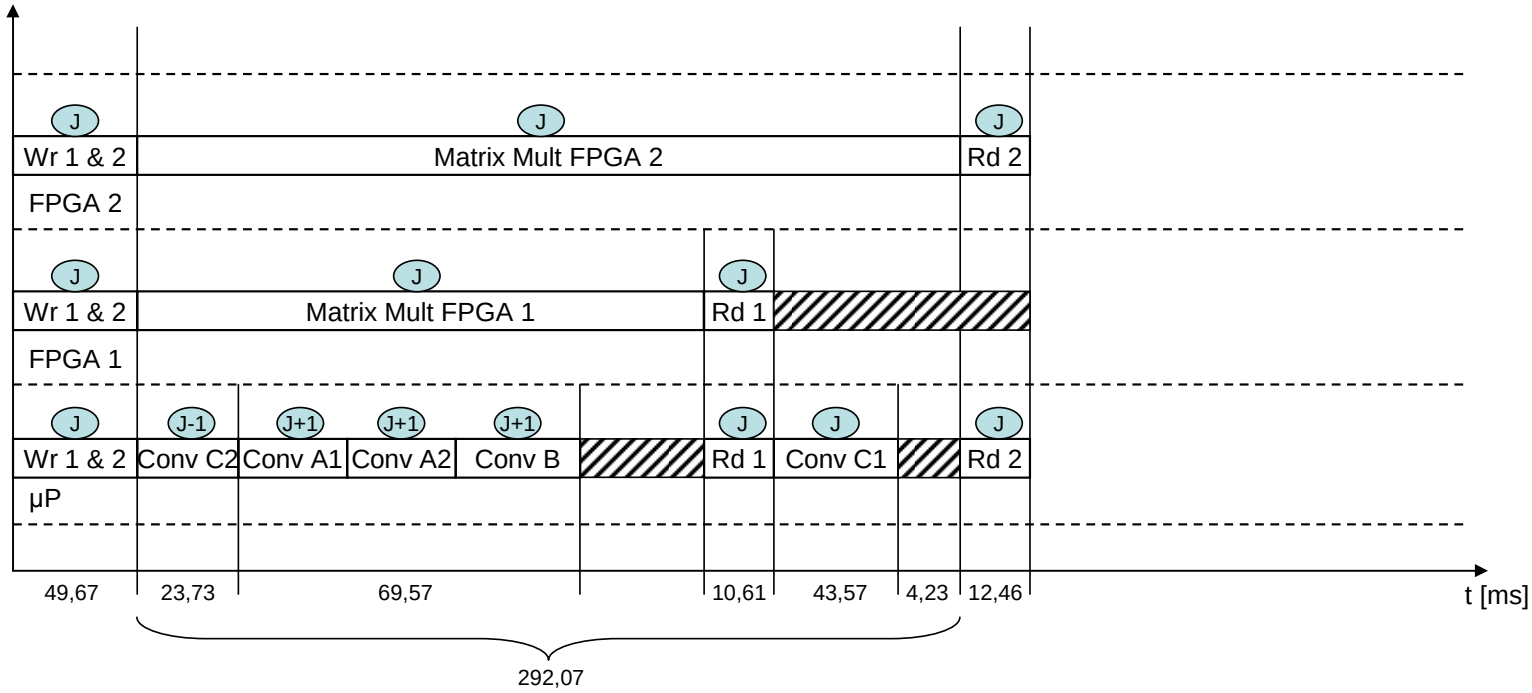
- Implementation of the "dgemm" function:
  - Interface: compliant with BLAS Library standard
  - Functionality: making usage of the FPGA board
- Performed operations :
  - Data conversions and data reorganisations
  - Data transfer management
  - Operation execution management

Outline

- Introduction
- Theoretical Study
- Prototype Implementation
- Measured Results
- Limitation Analysis
- Conclusion

Pipelined Chronogram

- Performances measured at **software level**



→Total: 374,06 ms

→Measured performances: 5,35 Gflops/s (2 FPGA)

ATLAS Software Implementation

- ATLAS = « Automatically Tuned Linear Algebra Software »
  - Optimised implementation of BLAS library
  - Measured performances (dgemm) :

    3,34 Gflops/s

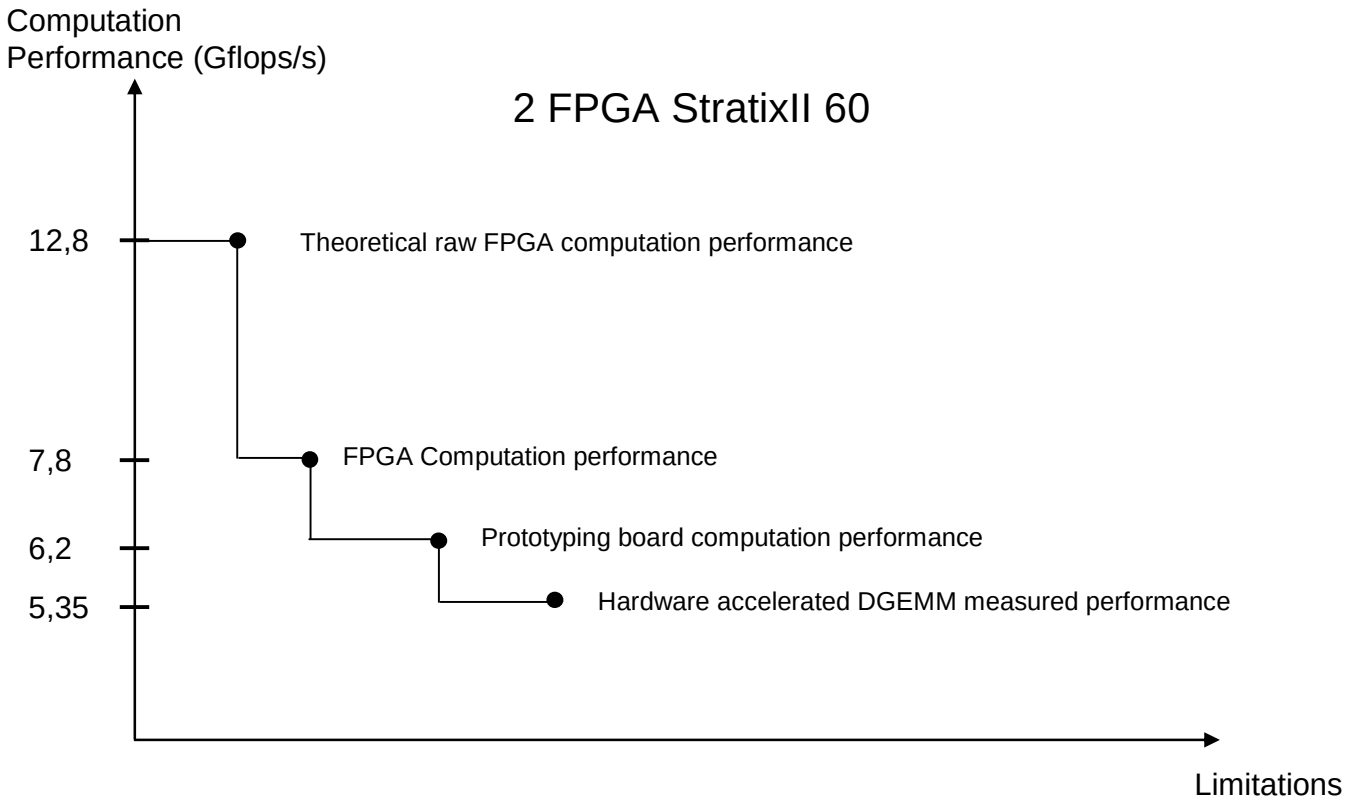
→Our implementation of dgemm:

60% higher performance

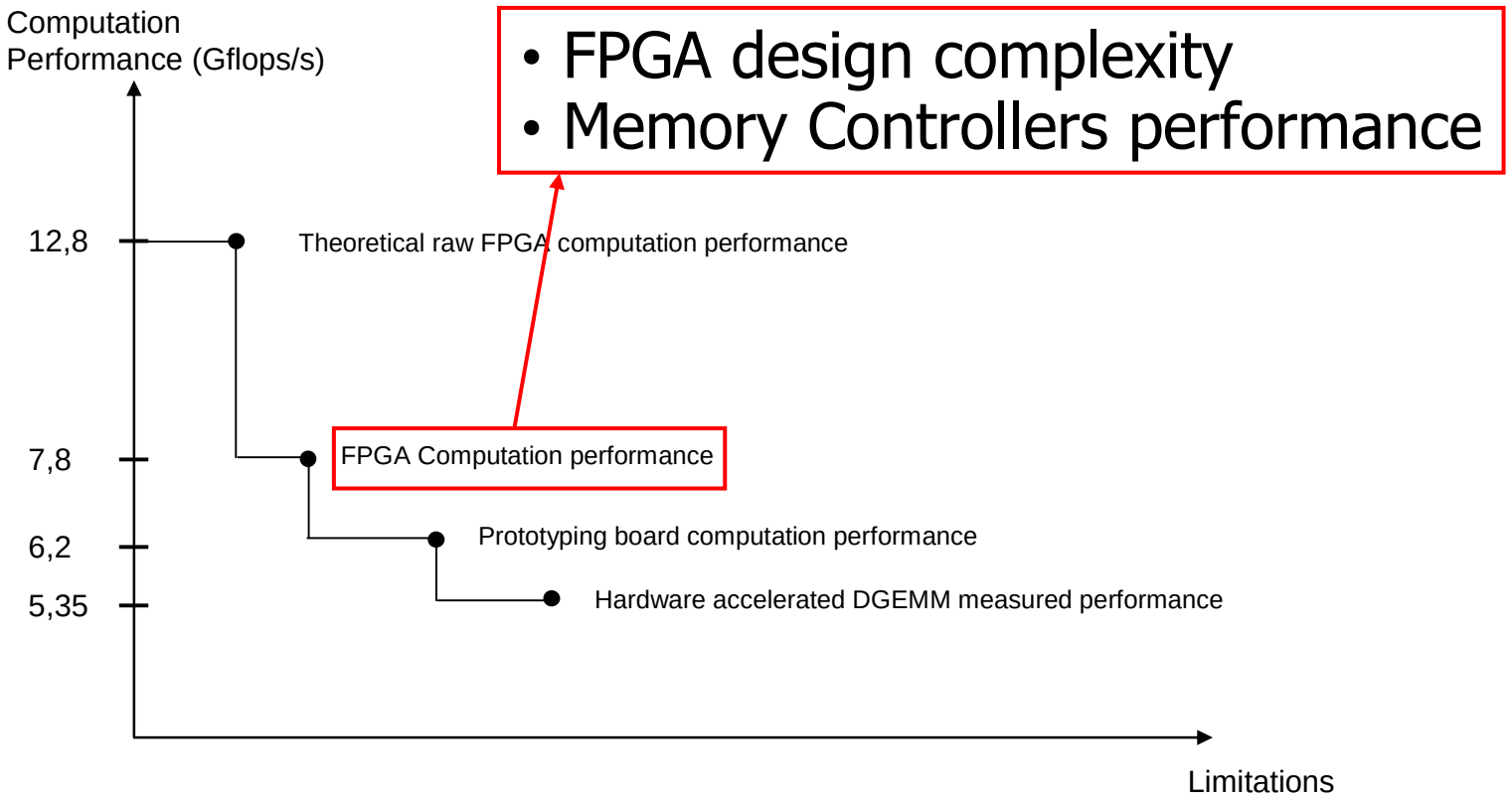
Outline

- Introduction
- Theoretical Study
- Prototype Implementation
- Measured Results
- Limitation Analysis
- Conclusion

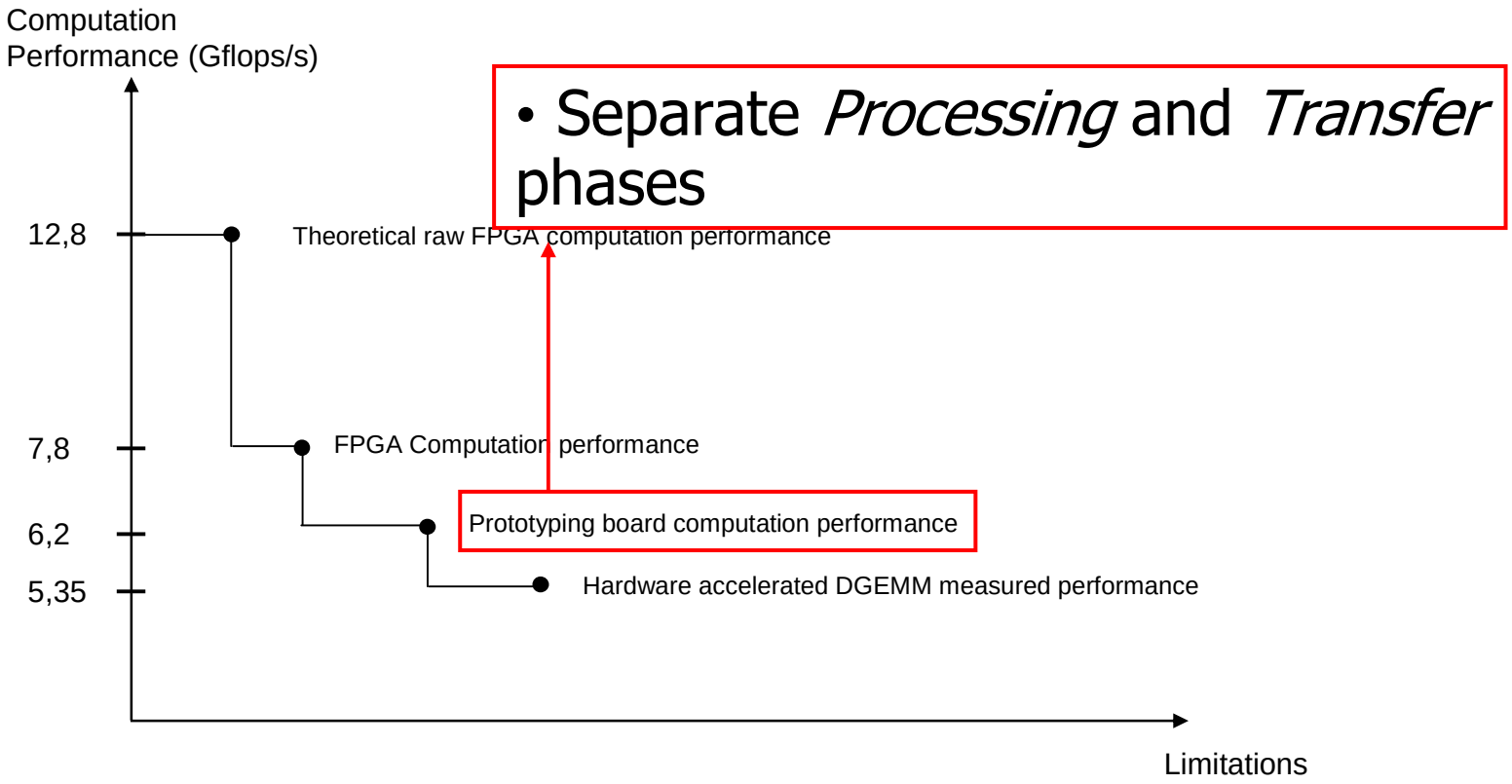
Performance Vs Limitations



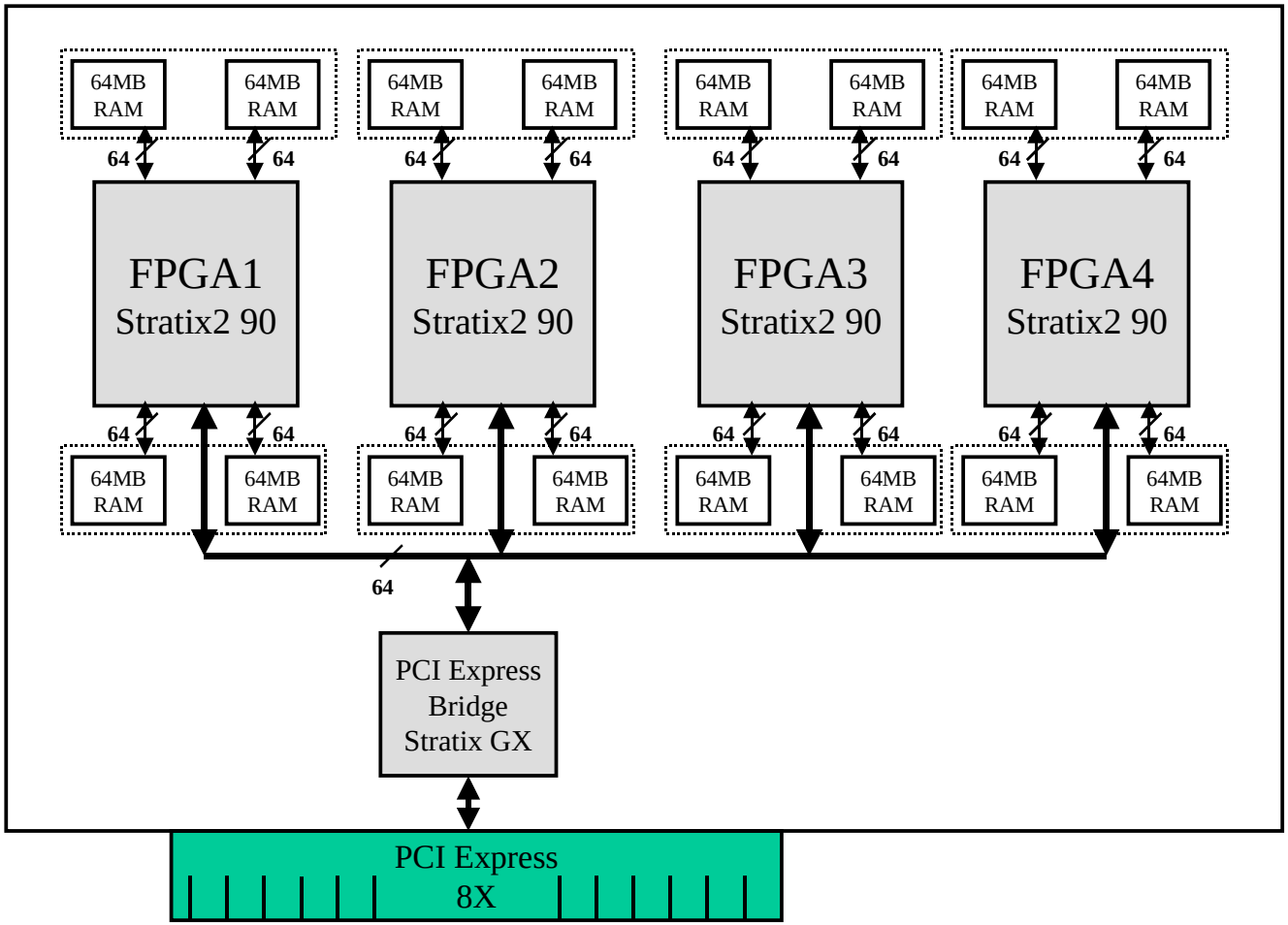
Performance Vs Limitations



Performance Vs Limitations



Adapted FPGA board



Outline

- Introduction
- Theoretical Study
- Prototype Implementation
- Measured Results
- Limitation Analysis
- Conclusion

Conclusion

- FPGAs deliver higher computation performance for the BLAS dgemm operation
  - 7,8 Gflops/s (2 FPGA StraticII 60)
- Main limitations come from external sources
  - Data transfers between the host computer and the FPGA board
  - Software data pre/post processing