Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

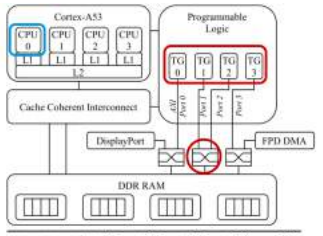
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive f embedded applications, but all face a similar problem

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H^{TG}_{01}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference

Algorithm 1: `stride` with intensity control.

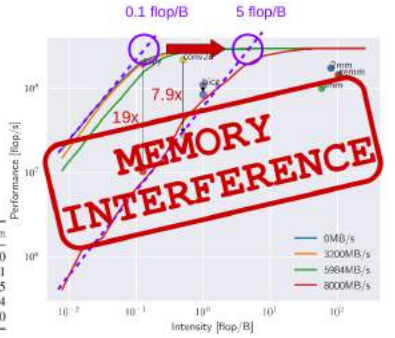
Data: vectors `X,Y` of length `n`, and a scalar `k`

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3 |   for j=0;j<k;j++ do
4 |   |   Y[i]+=X[i];
5 |   end
6 end
```

QUANTITATIVE IMPACT

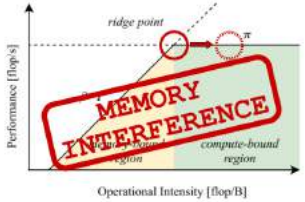
- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.22	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



NEW INSIGHTS

- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

SUMMARY AND CONCLUSION

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x
  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, 2009.

[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

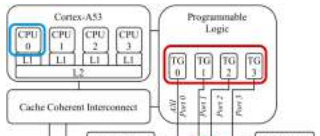
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive f embedded applications, but all face a similar problem

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark
- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference



Injected BW	H^{TG}_{01}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

Algorithm 1: `stride` with intensity control.

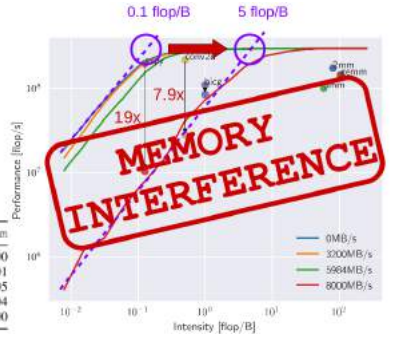
Data: vectors X,Y of length n, and a scalar k

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3 |   for j=0;j<k;j++ do
4 |   |   Y[i]+=X[i];
5 |   end
6 end
```

QUANTITATIVE IMPACT

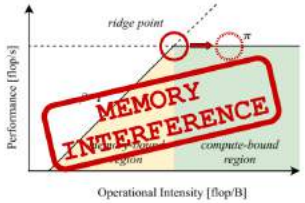
- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.22	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



NEW INSIGHTS

- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

SUMMARY AND CONCLUSION

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x
  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, 2009.

[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

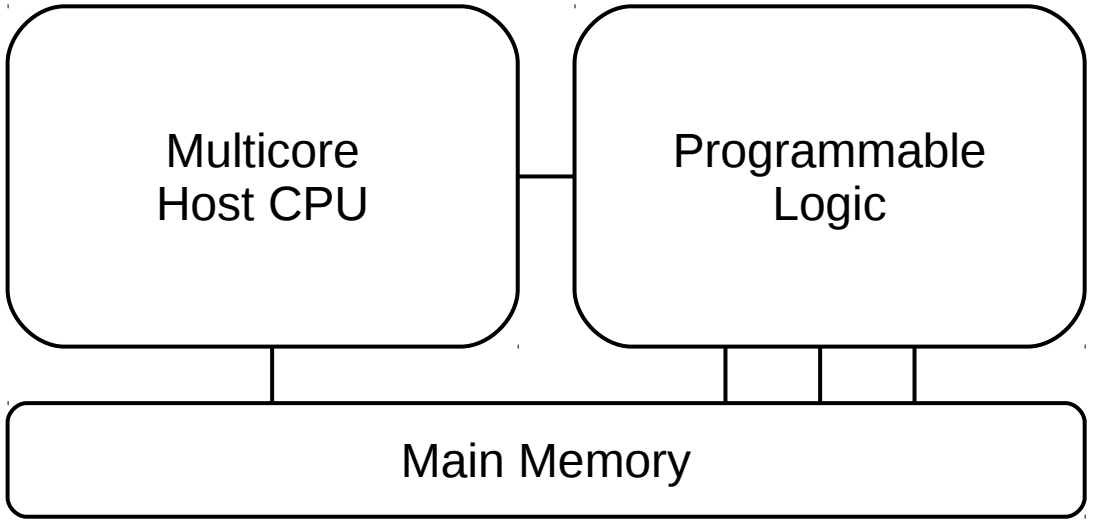
[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

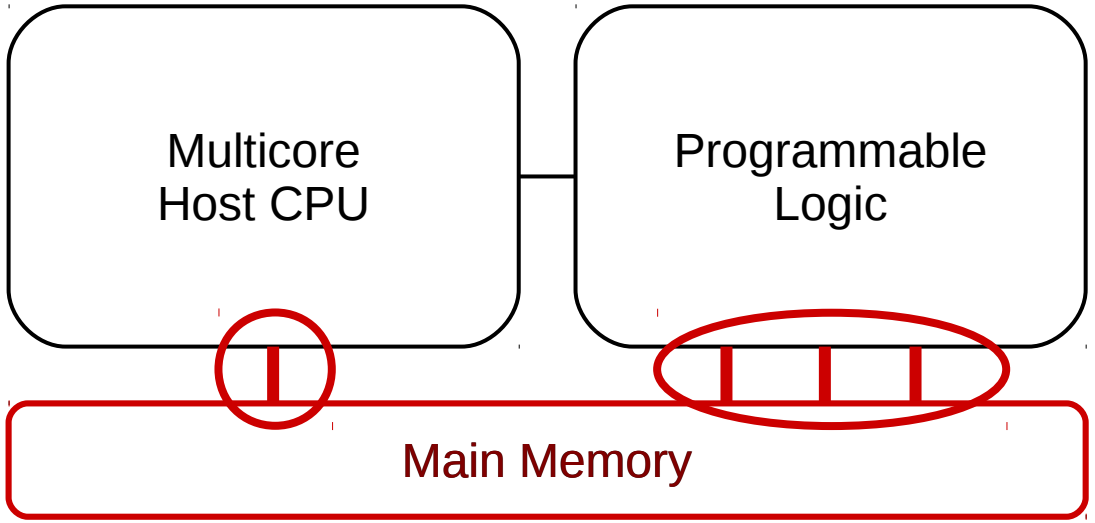
Background & Motivation

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem



Background & Motivation

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem



Background & Motivation

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

Background & Motivation

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

Background & Motivation

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

Background & Motivation

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

Background & Motivation

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

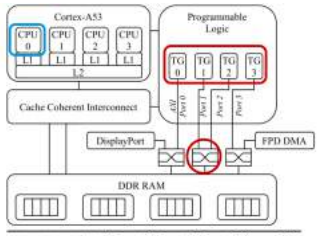
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H^{TG}_{01}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference

Algorithm 1: `stride` with intensity control.

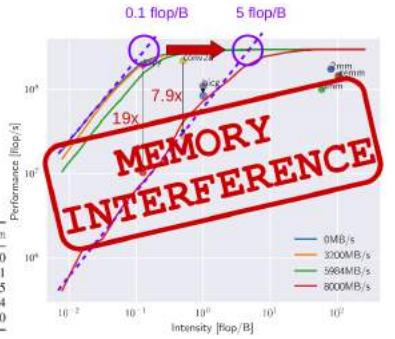
Data: vectors `X,Y` of length `n`, and a scalar `k`

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3 |   for j=0;j<k;j++ do
4 |   |   Y[i]+=X[i];
5 |   end
6 end
```

QUANTITATIVE IMPACT

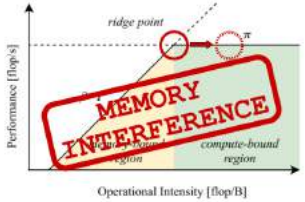
- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.22	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



NEW INSIGHTS

- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, 2009.

[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

SUMMARY AND CONCLUSION

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x
  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

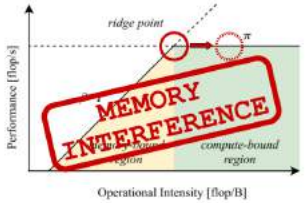
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

NEW INSIGHTS

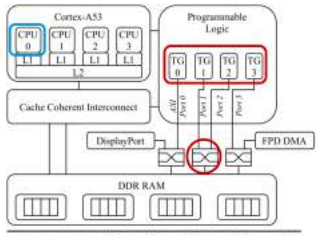
- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H^{TG}_{01}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference

Algorithm 1: `stride` with intensity control.

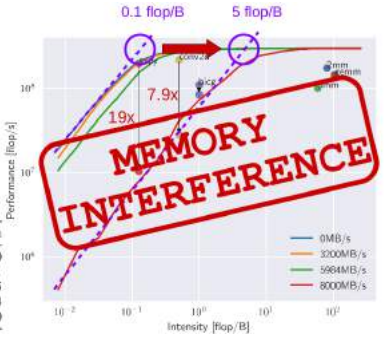
Data: vectors `X,Y` of length `n`, and a scalar `k`

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3 |   for j=0;j<k;j++ do
4 |   |   Y[i]+=X[i];
5 |   end
6 end
```

QUANTITATIVE IMPACT

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.22	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



SUMMARY AND CONCLUSION

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x
  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, 2009.

[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

New Insights

New Insights

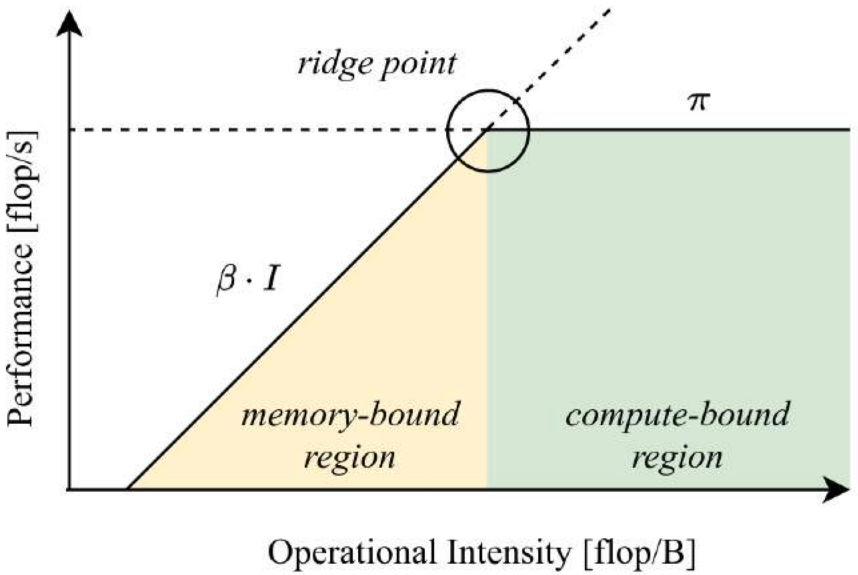
- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+

New Insights

- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark

New Insights

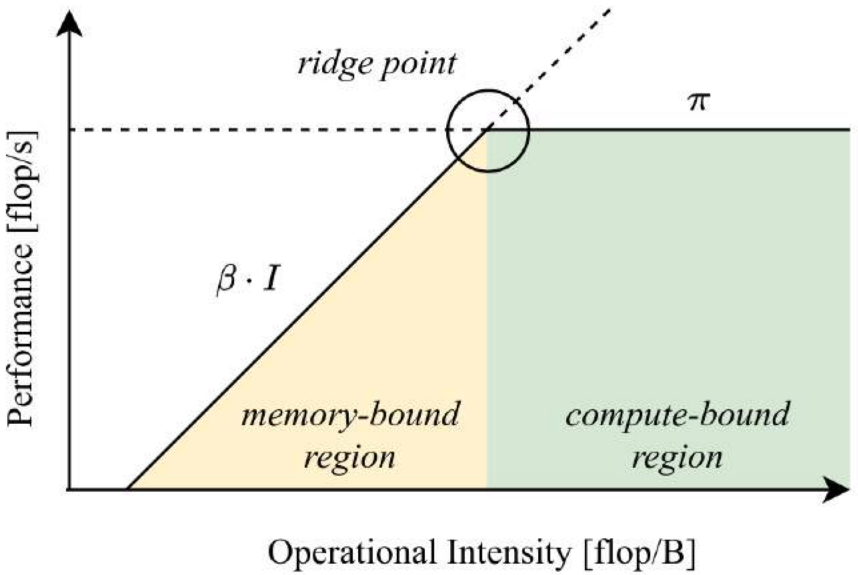
- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]

New Insights

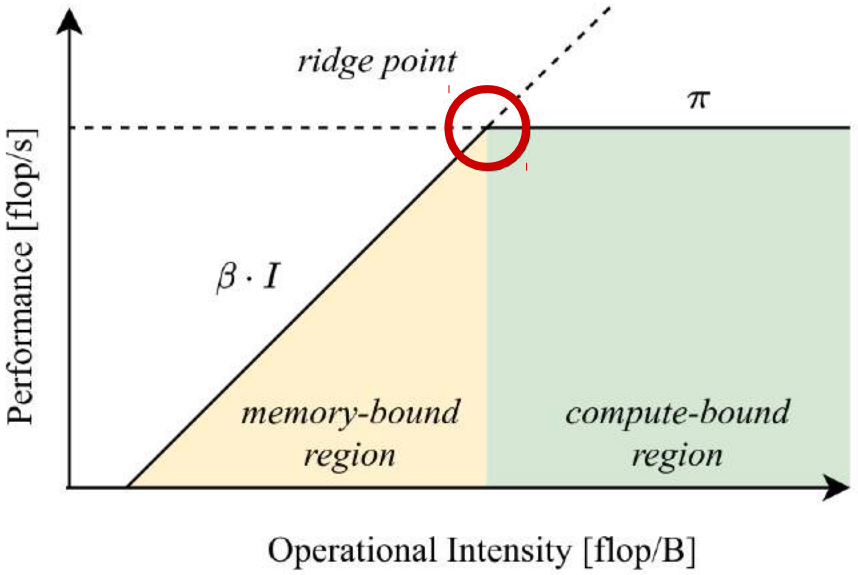
- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

New Insights

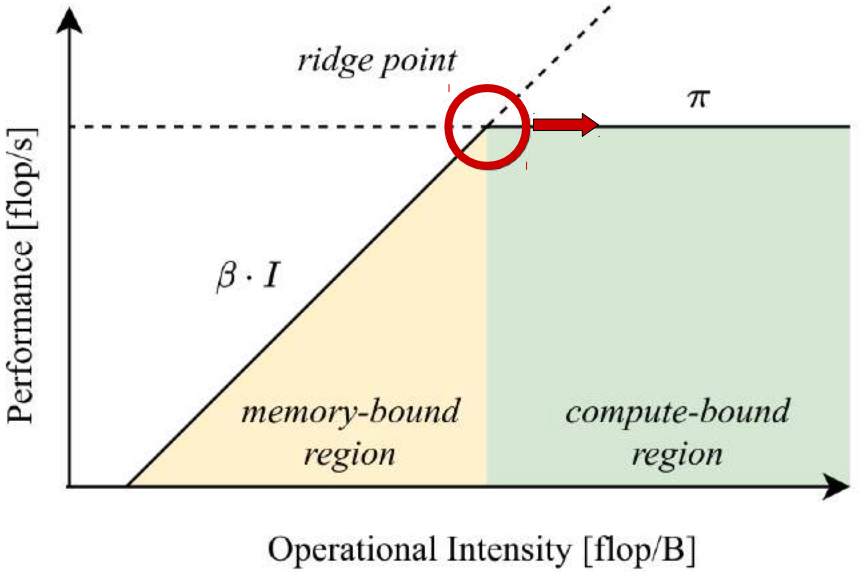
- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

New Insights

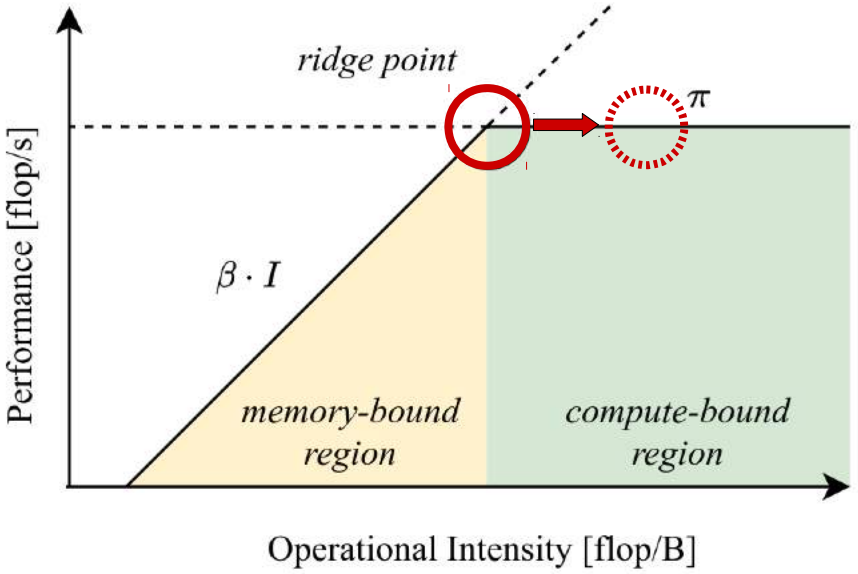
- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

New Insights

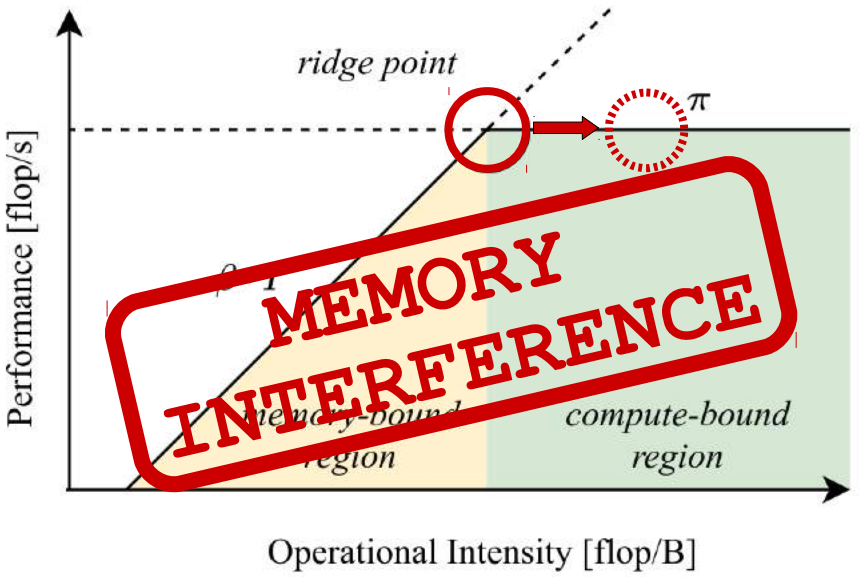
- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

New Insights

- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

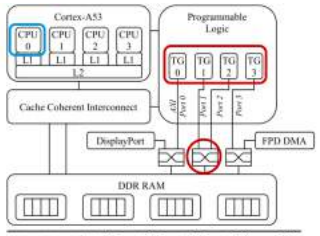
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H^{TG}_{01}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference

Algorithm 1: `stride` with intensity control.

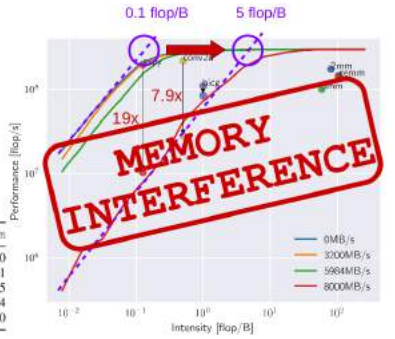
Data: vectors `X,Y` of length `n`, and a scalar `k`

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3 |   for j=0;j<k;j++ do
4 |   |   Y[i]+=X[i];
5 |   end
6 end
```

QUANTITATIVE IMPACT

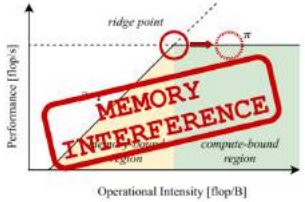
- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.22	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



NEW INSIGHTS

- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, 2009.

[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

SUMMARY AND CONCLUSION

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x
  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

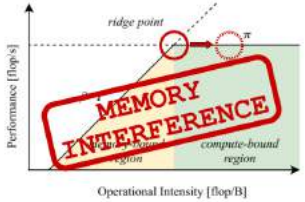
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive fo
embedded applications, but all face a similar problem

NEW INSIGHTS

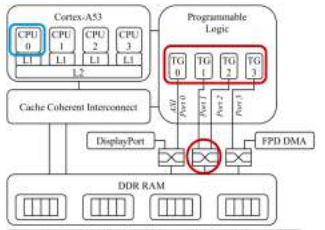
- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H^{TG}_{01}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference

Algorithm 1: `stride` with intensity control.

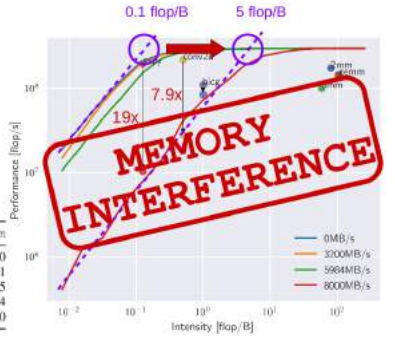
Data: vectors `X,Y` of length `n`, and a scalar `k`

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3 |   for j=0;j<k;j++ do
4 |   |   Y[i]+=X[i];
5 |   end
6 end
```

QUANTITATIVE IMPACT

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.22	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



SUMMARY AND CONCLUSION

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x
  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, 2009.

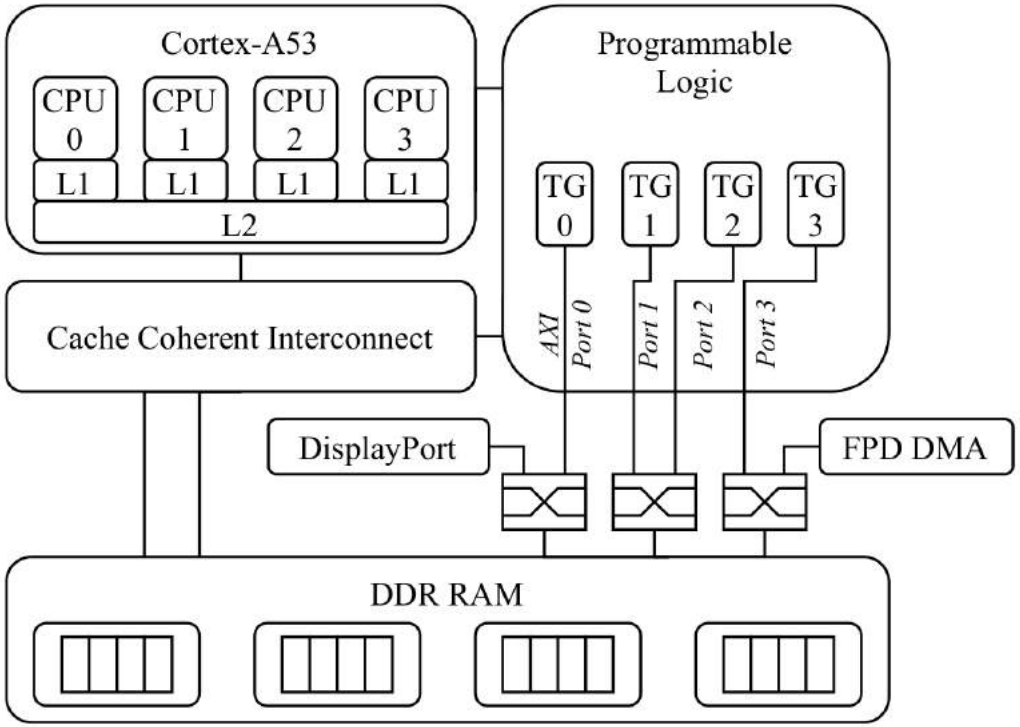
[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

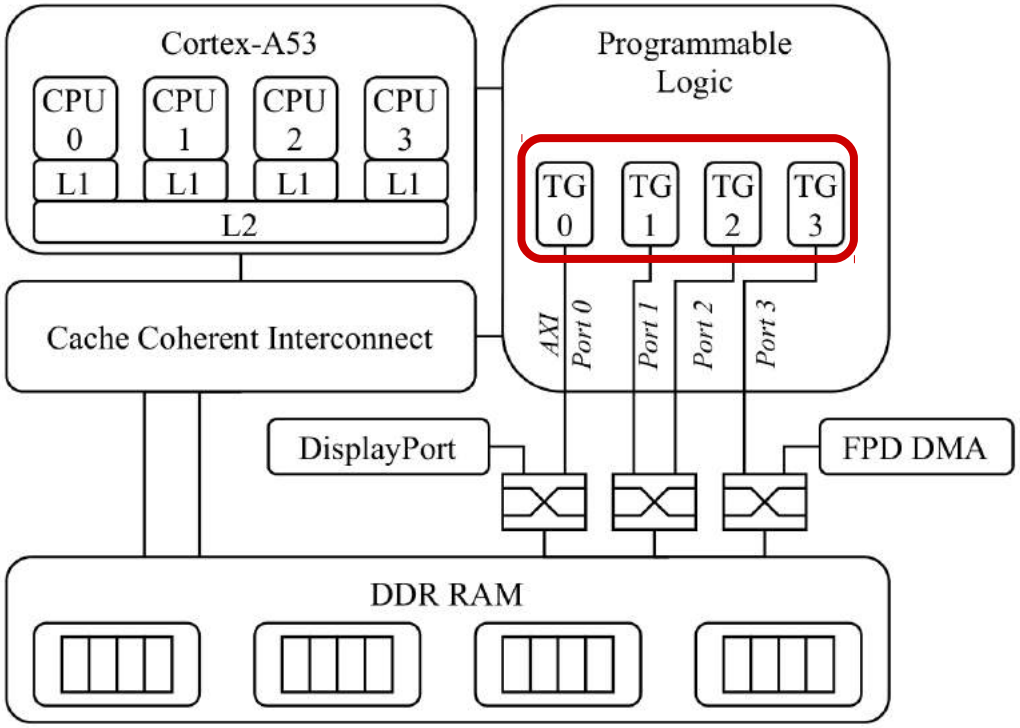
[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

Description I



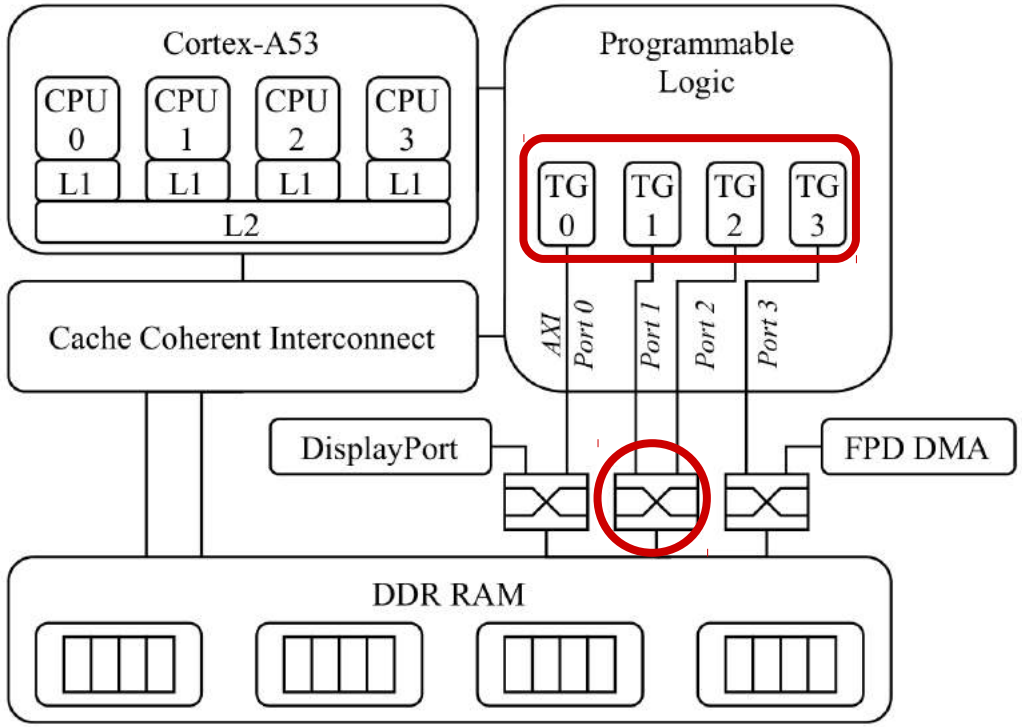
Description I

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM



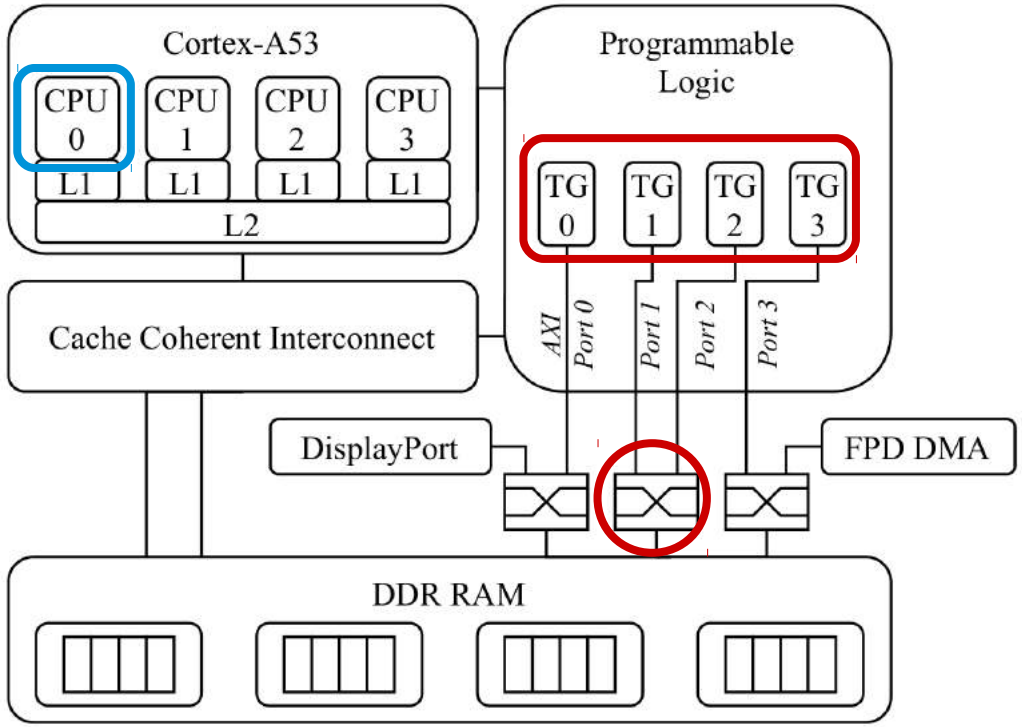
Description I

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM



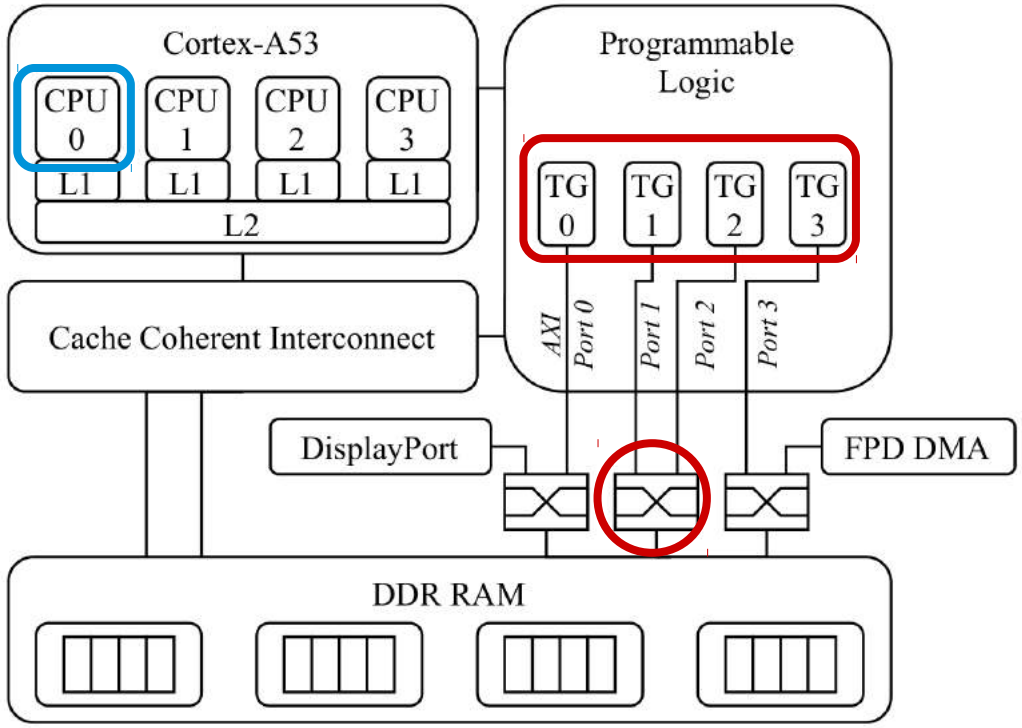
Description I

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Description I

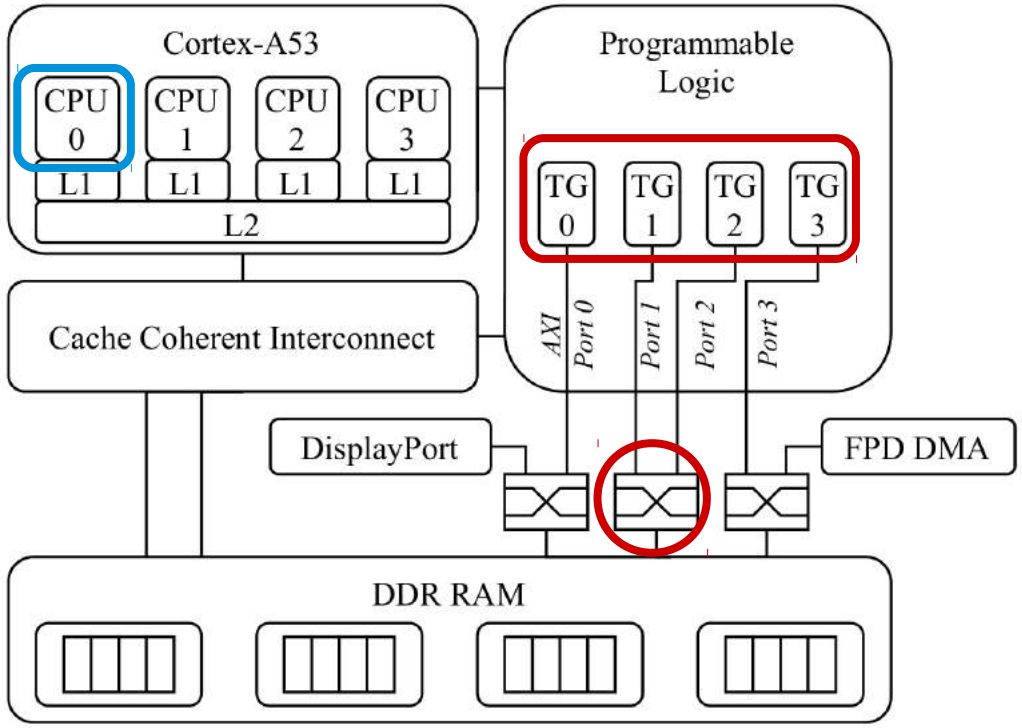
- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW
0 MB/s
320 MB/s
3200 MB/s
5984 MB/s
8000 MB/s

Description I

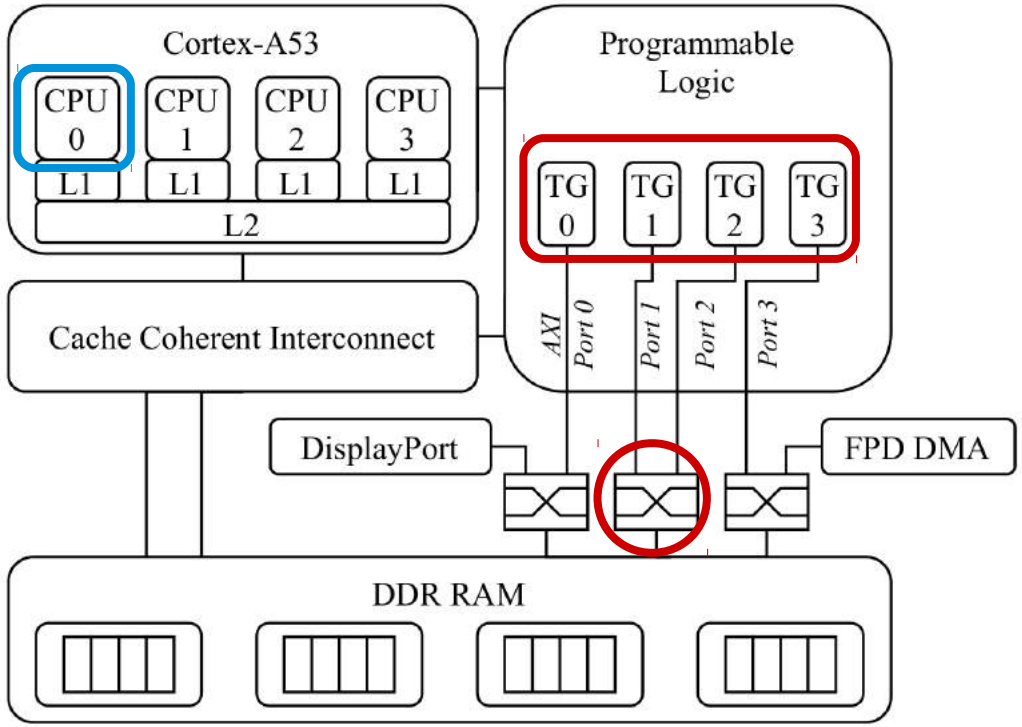
- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H_{01}^{TG}
0 MB/s	1.00
320 MB/s	1.17
3200 MB/s	1.26
5984 MB/s	2.51
8000 MB/s	4.37

Description I

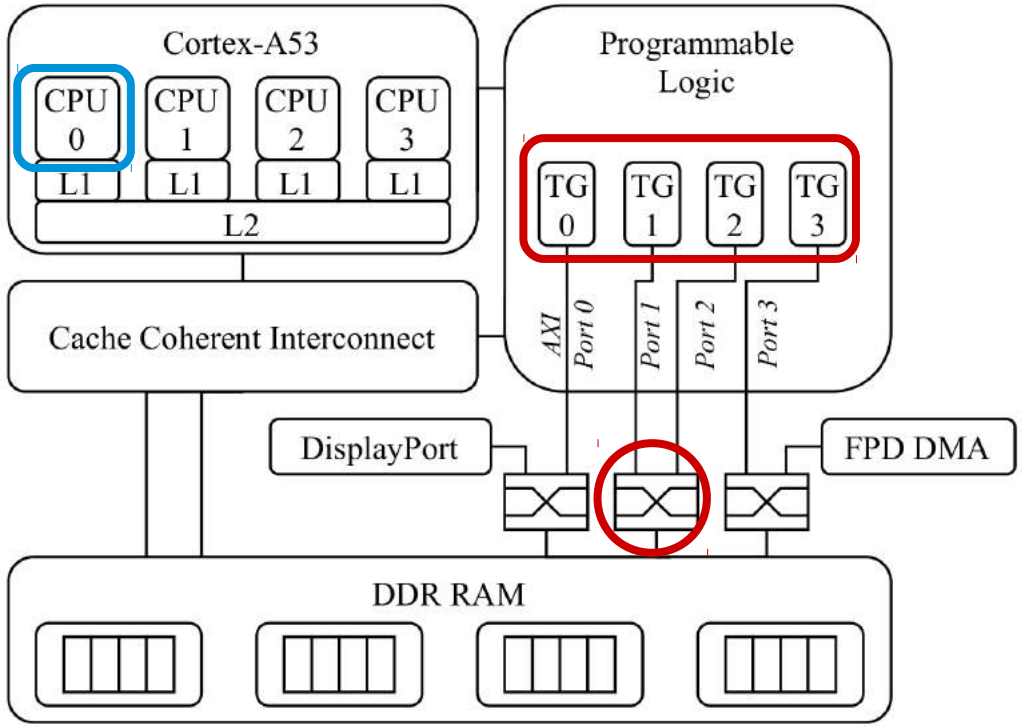
- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H_{01}^{TG}	H_{12}^{TG}
0 MB/s	1.00	1.00
320 MB/s	1.17	1.12
3200 MB/s	1.26	1.29
5984 MB/s	2.51	3.01
8000 MB/s	4.37	6.51

Description I

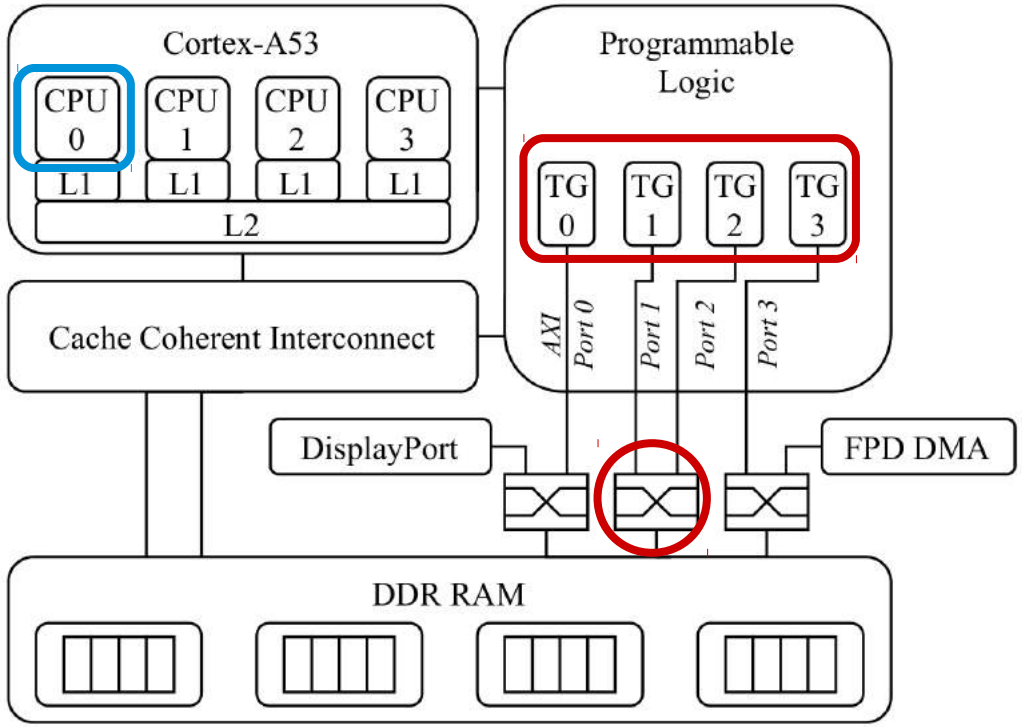
- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H_{01}^{TG}	H_{12}^{TG}	H_{013}^{TG}	H_{012}^{TG}	H_{0123}^{TG}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.99

Description I

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H^{TG}_{01}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.99

Description II

- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference

Algorithm 1: `stride` with intensity control.

Data: vectors `X,Y` of length `n`, and a scalar `k`

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3 |   for j=0;j<k;j++ do
4 |   |   Y[i]+=X[i];
5 |   end
6 end
```

Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

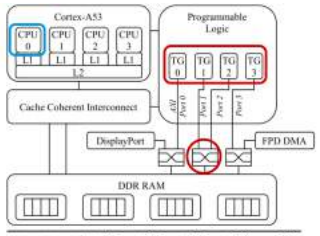
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H^{TG}_{01}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference

Algorithm 1: `stride` with intensity control.

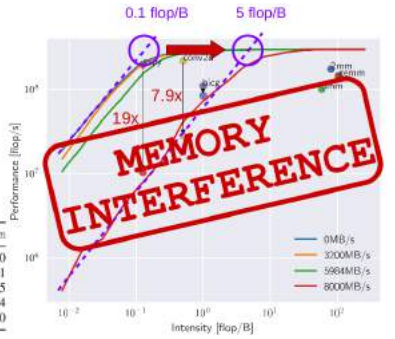
Data: vectors `X,Y` of length `n`, and a scalar `k`

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3 |   for j=0;j<k;j++ do
4 |   |   Y[i]+=X[i];
5 |   end
6 end
```

QUANTITATIVE IMPACT

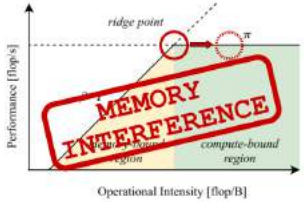
- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.22	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



NEW INSIGHTS

- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, 2009.

[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

SUMMARY AND CONCLUSION

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x
  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

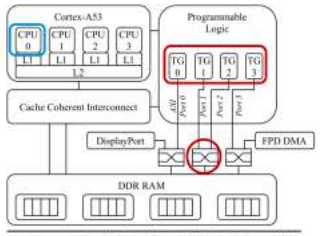
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H^{TG}_{00}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference

Algorithm 1: `stride` with intensity control.

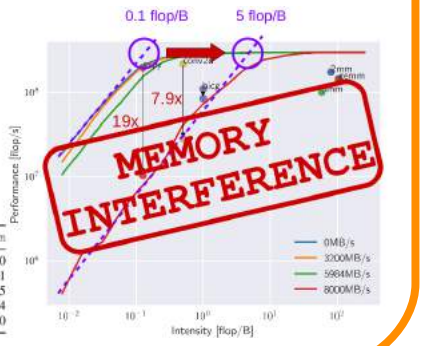
Data: vectors `X,Y` of length `n`, and a scalar `k`

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3 |   for j=0;j<k;j++ do
4 |   |   Y[i]+=X[i];
5 |   end
6 end
```

QUANTITATIVE IMPACT

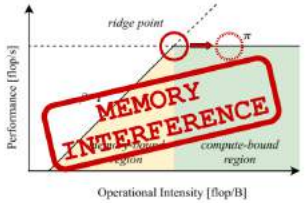
- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.22	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



NEW INSIGHTS

- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, 2009.

[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

SUMMARY AND CONCLUSION

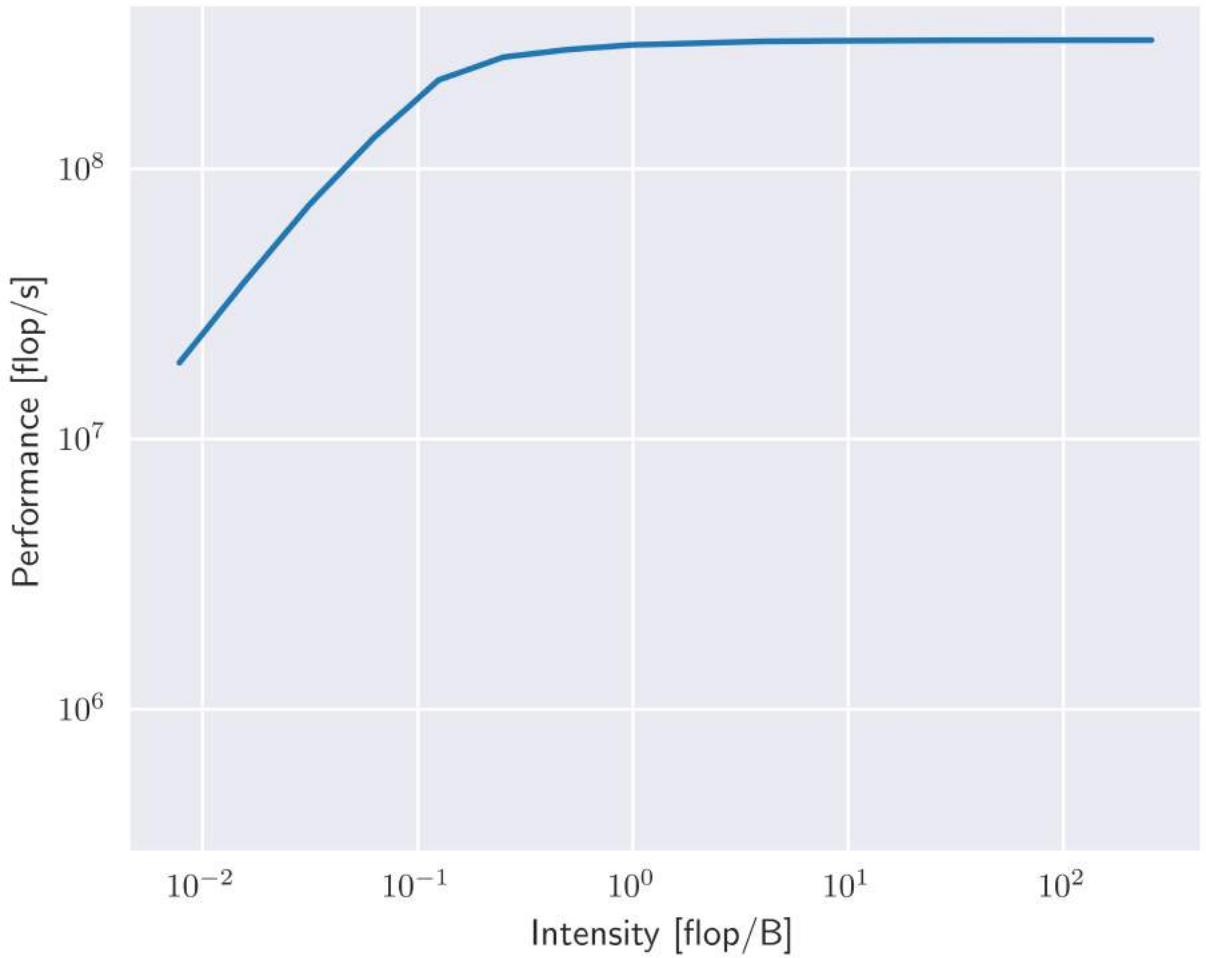
- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x
  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

Quantitative Impact

- Jitter:
  - Maximal deviation from median

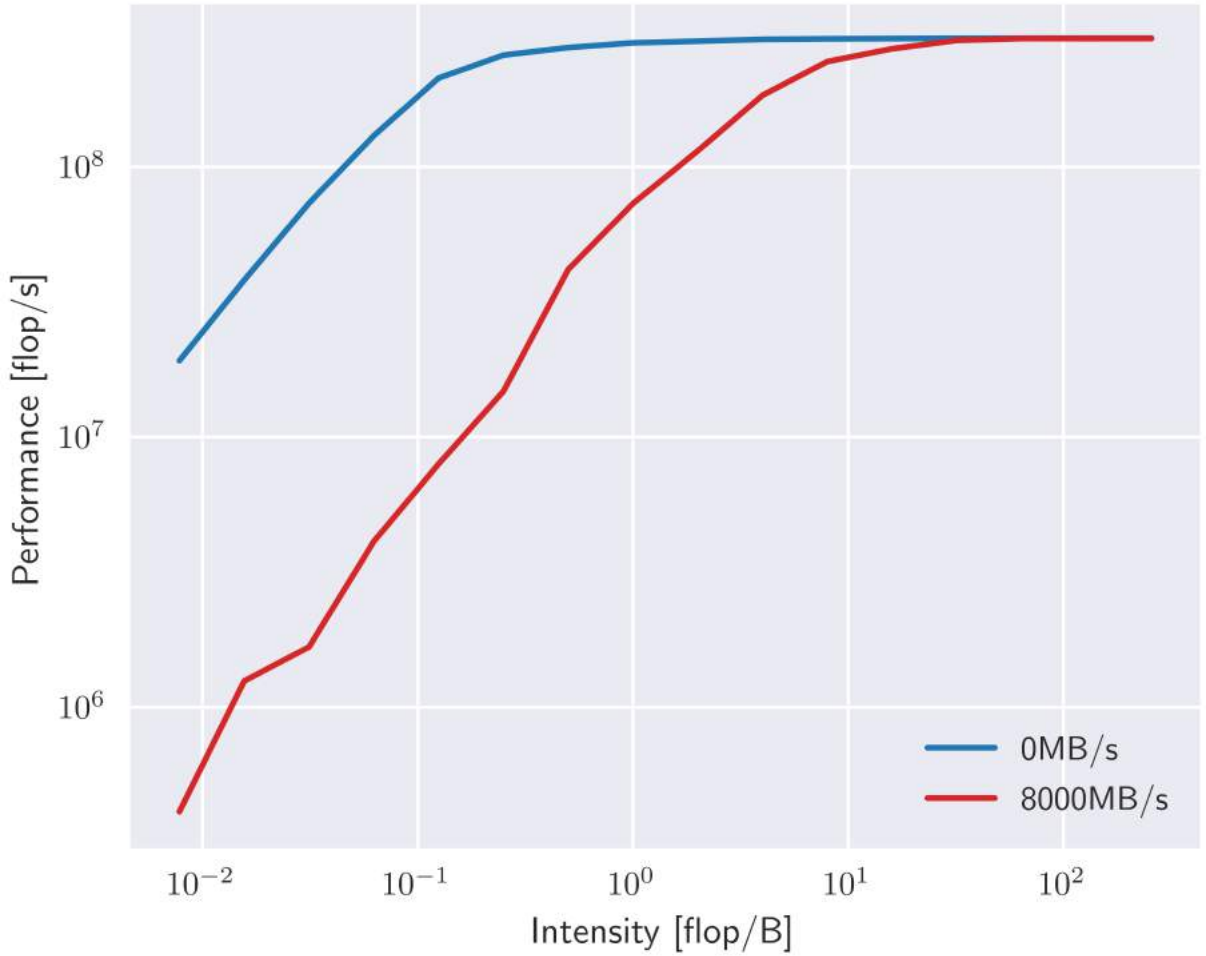
Quantitative Impact

- Jitter:
  - Maximal deviation from median



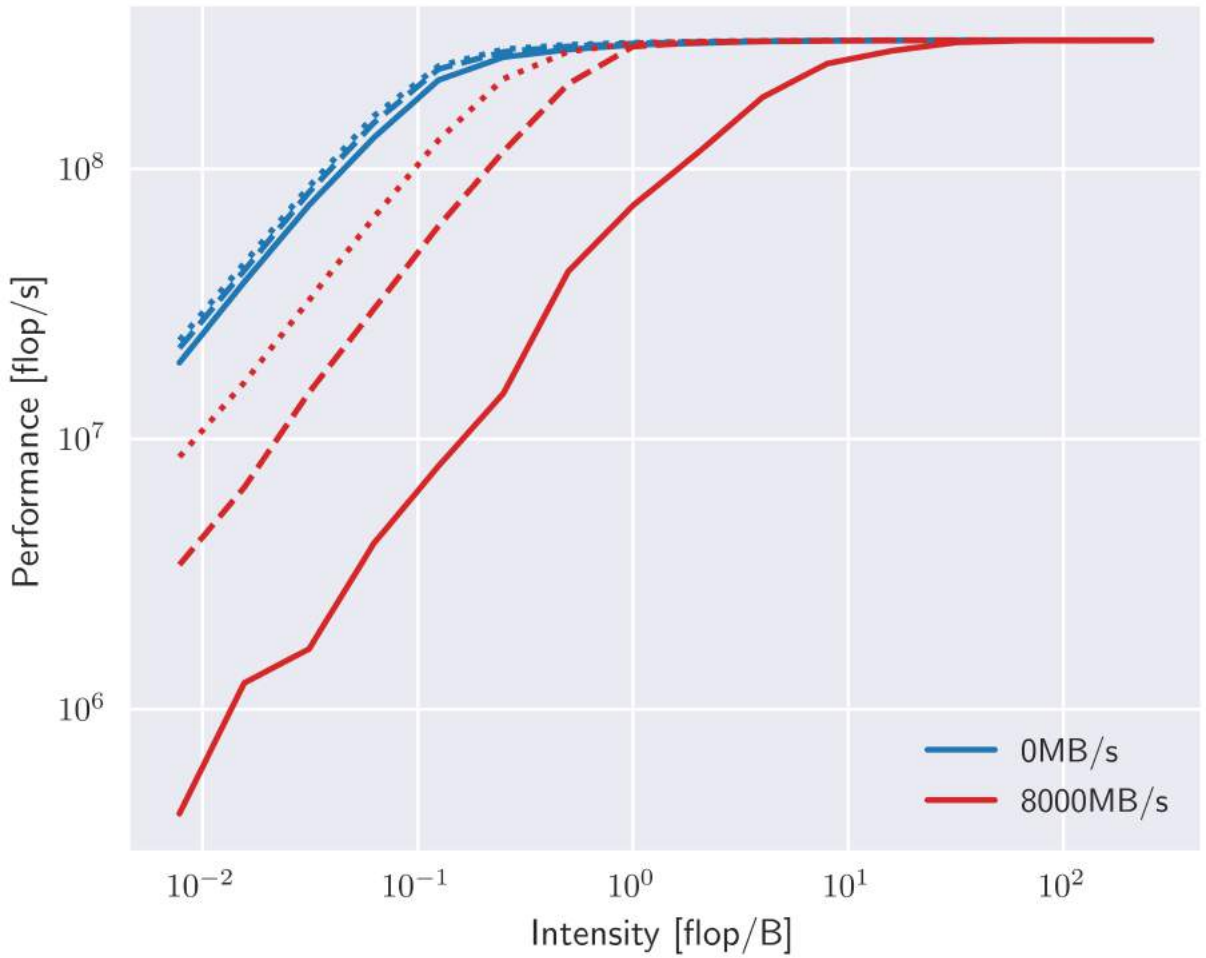
Quantitative Impact

- Jitter:
  - Maximal deviation from median



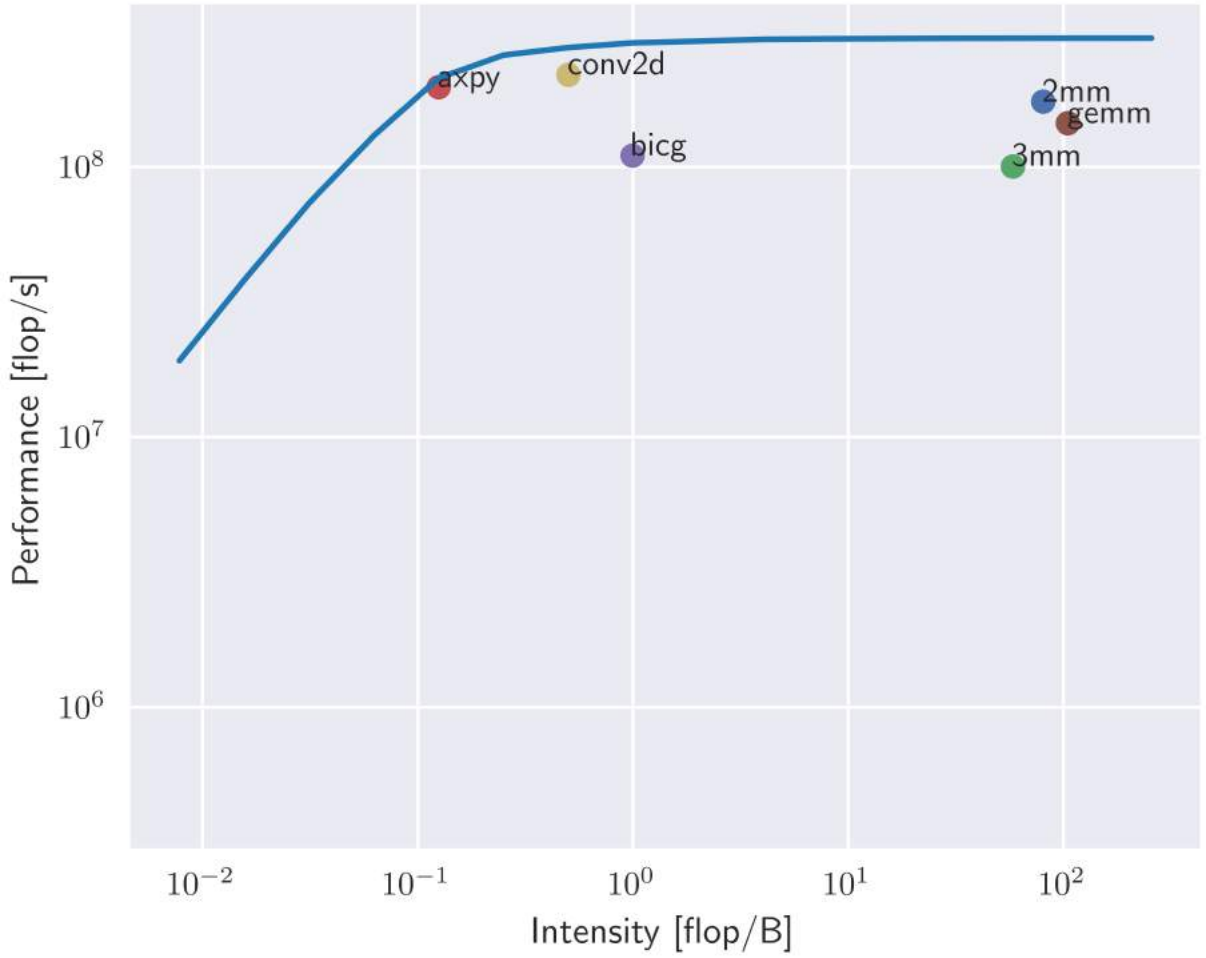
Quantitative Impact

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference



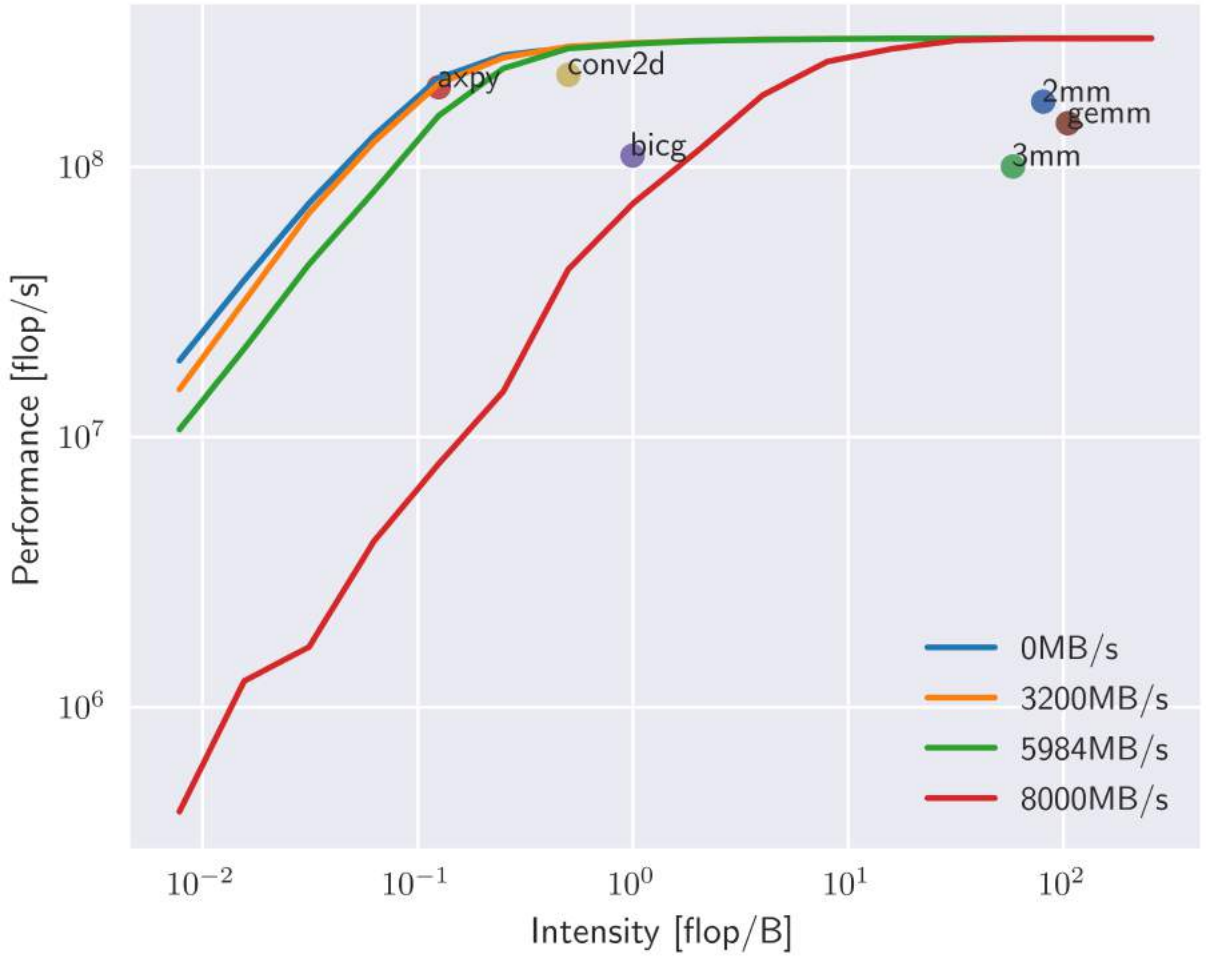
Quantitative Impact

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown



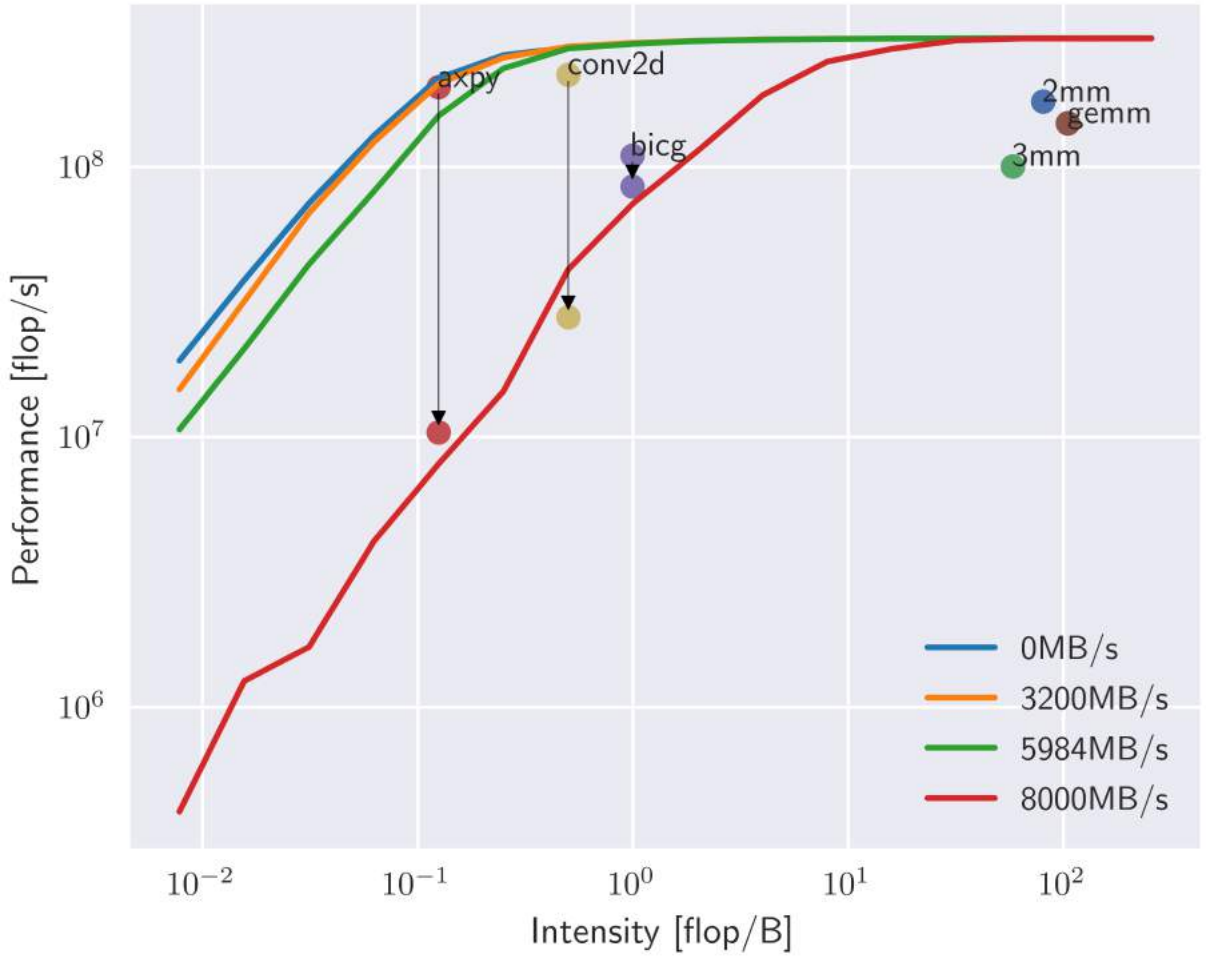
Quantitative Impact

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown



Quantitative Impact

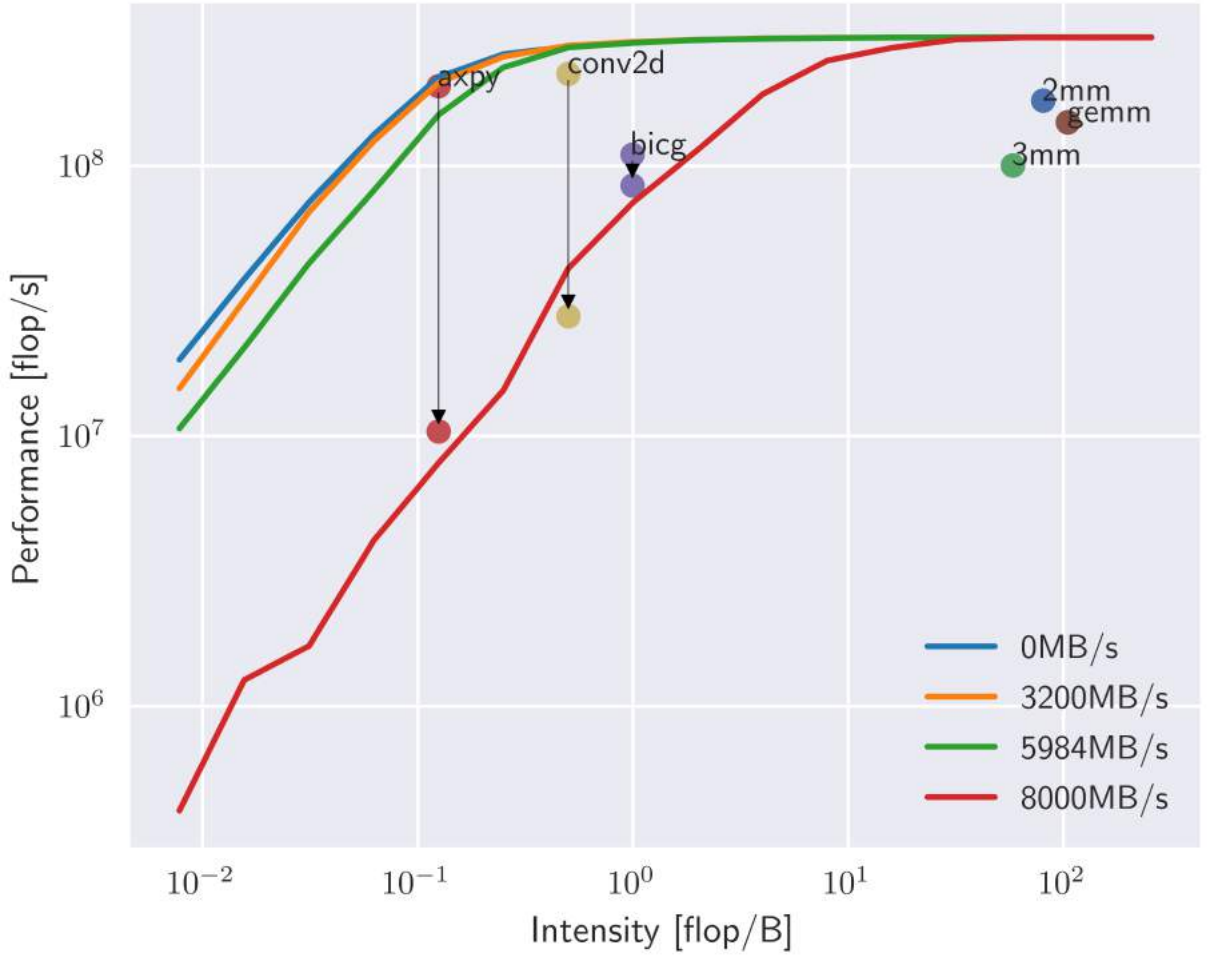
- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown



Quantitative Impact

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

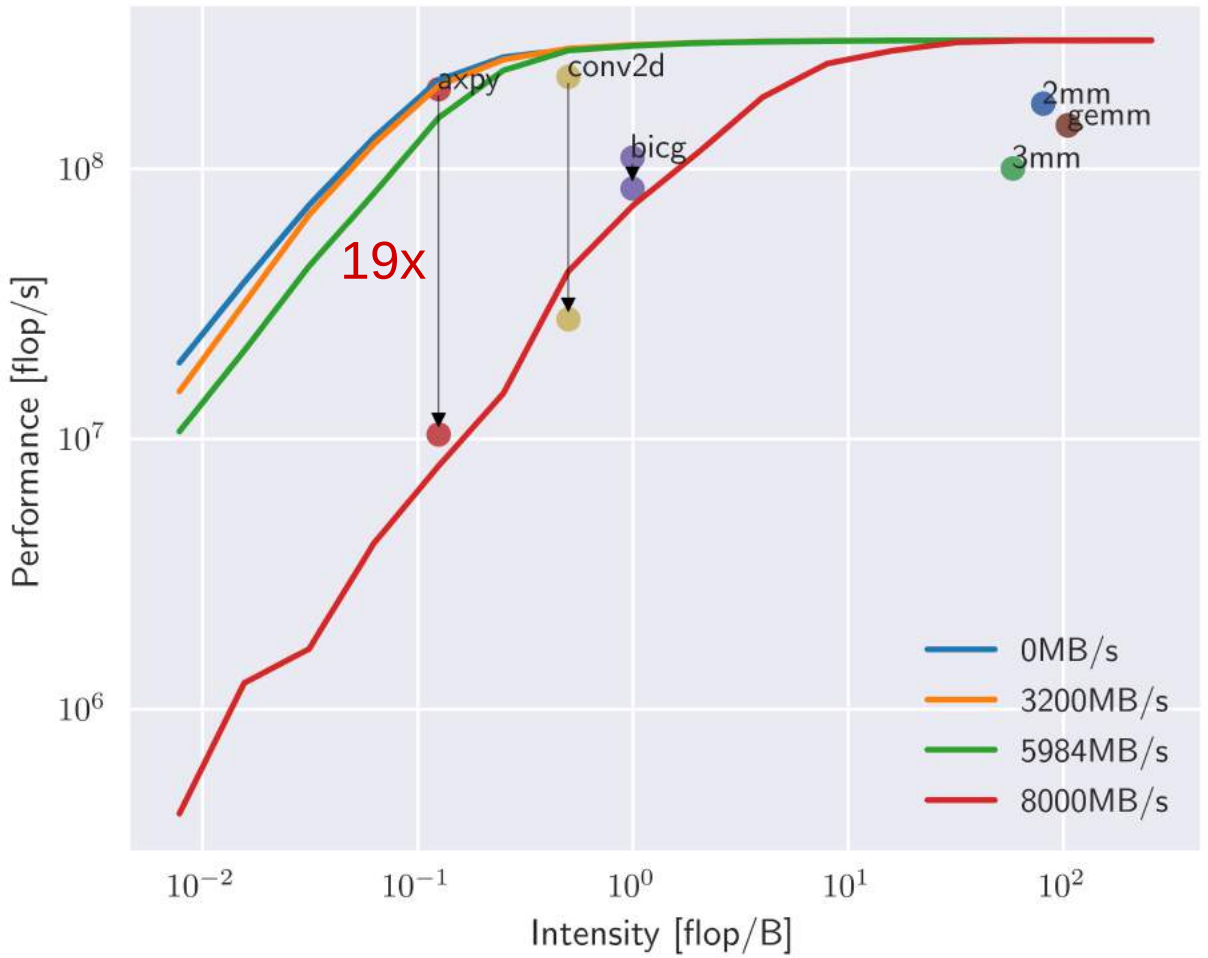
Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.27	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



Quantitative Impact

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

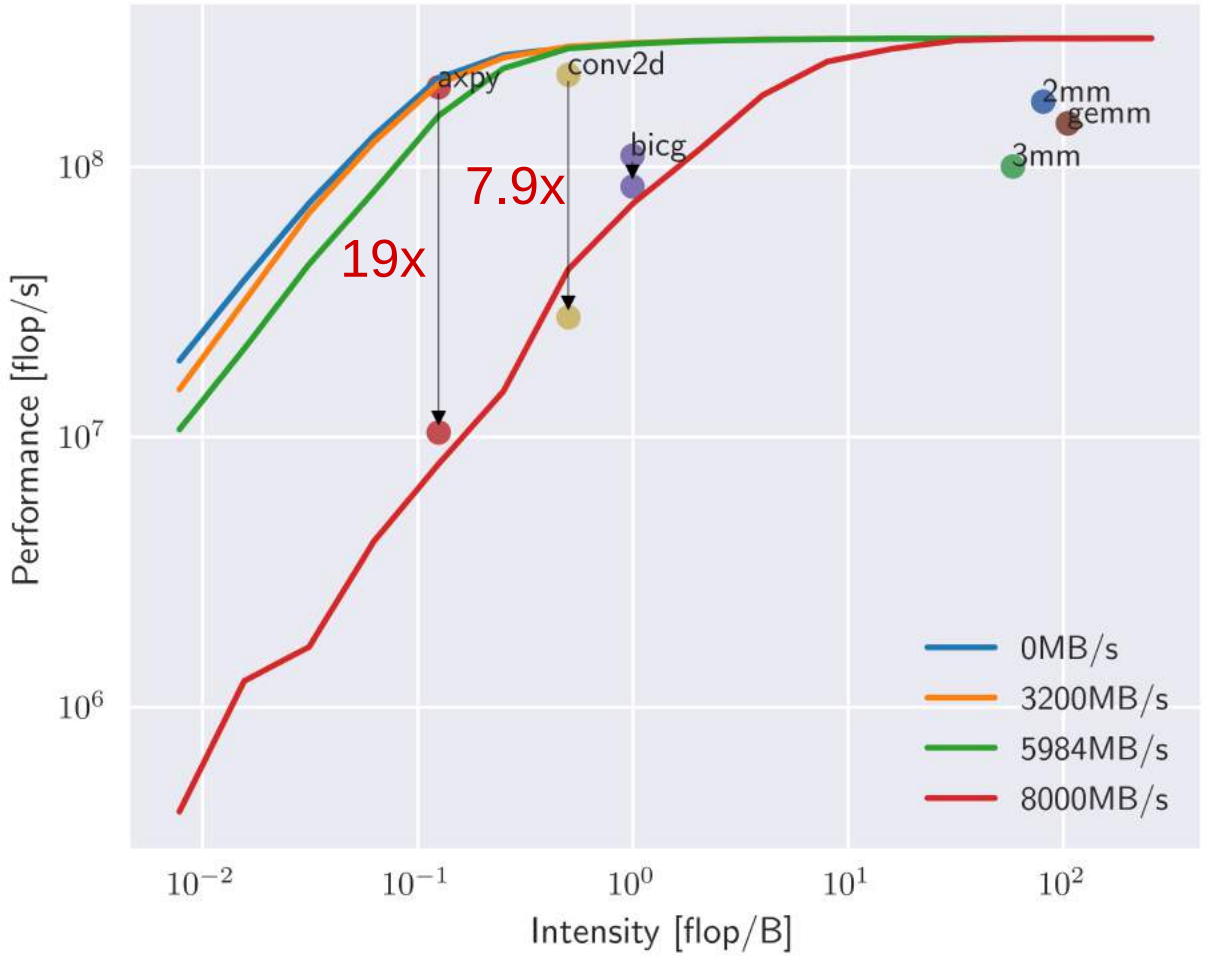
Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.27	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



Quantitative Impact

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

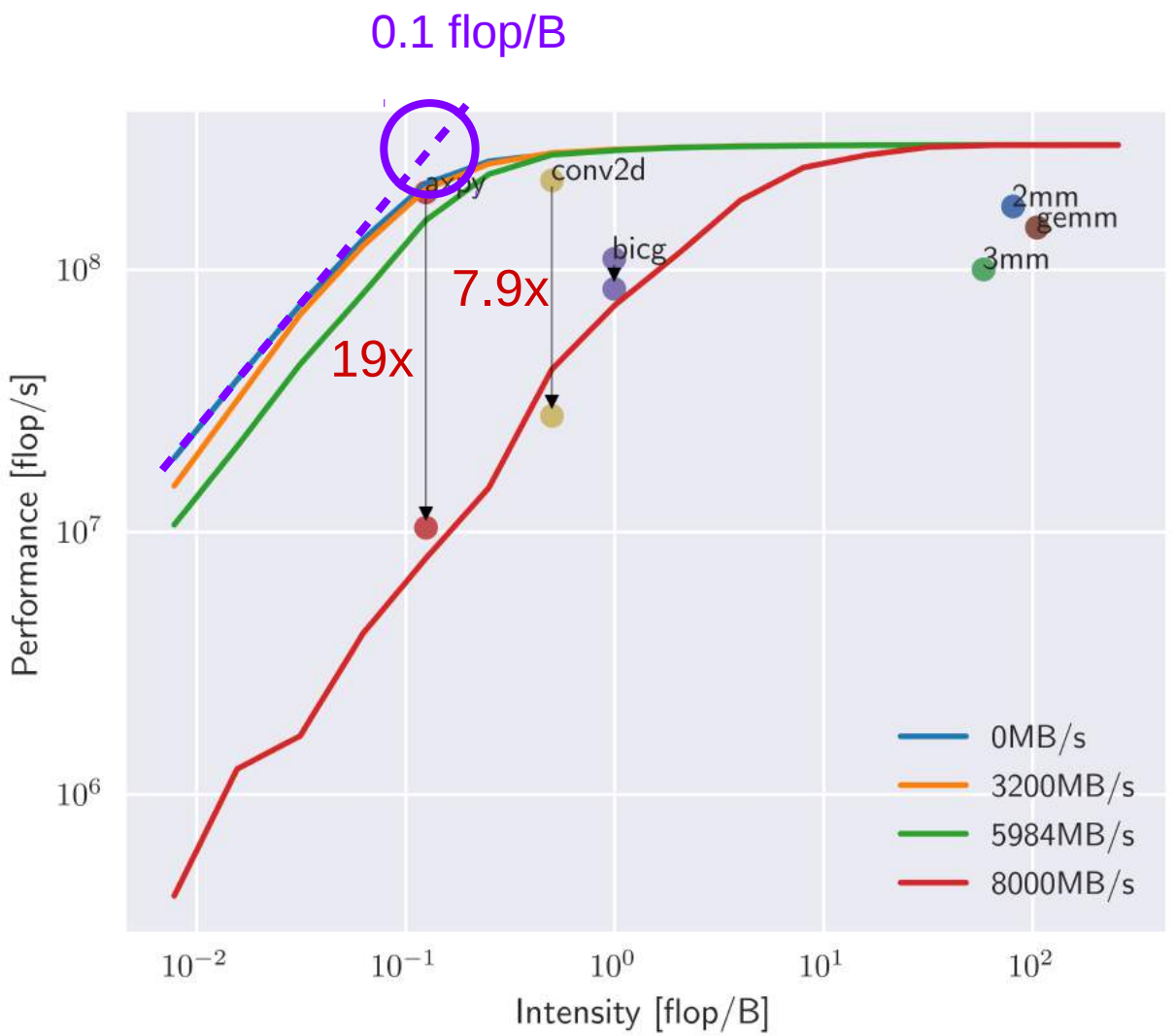
Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.27	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



Quantitative Impact

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

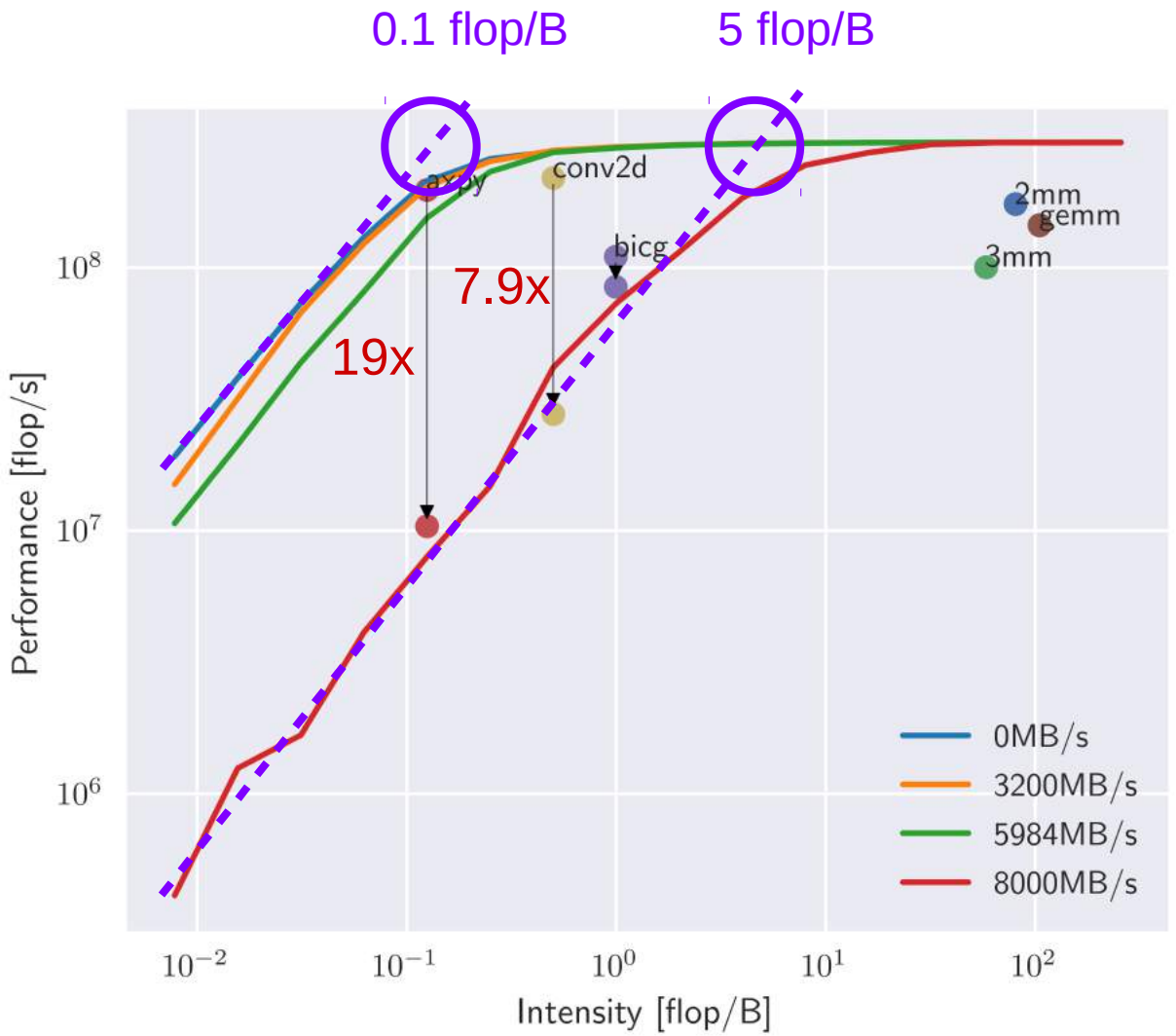
Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.27	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



Quantitative Impact

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

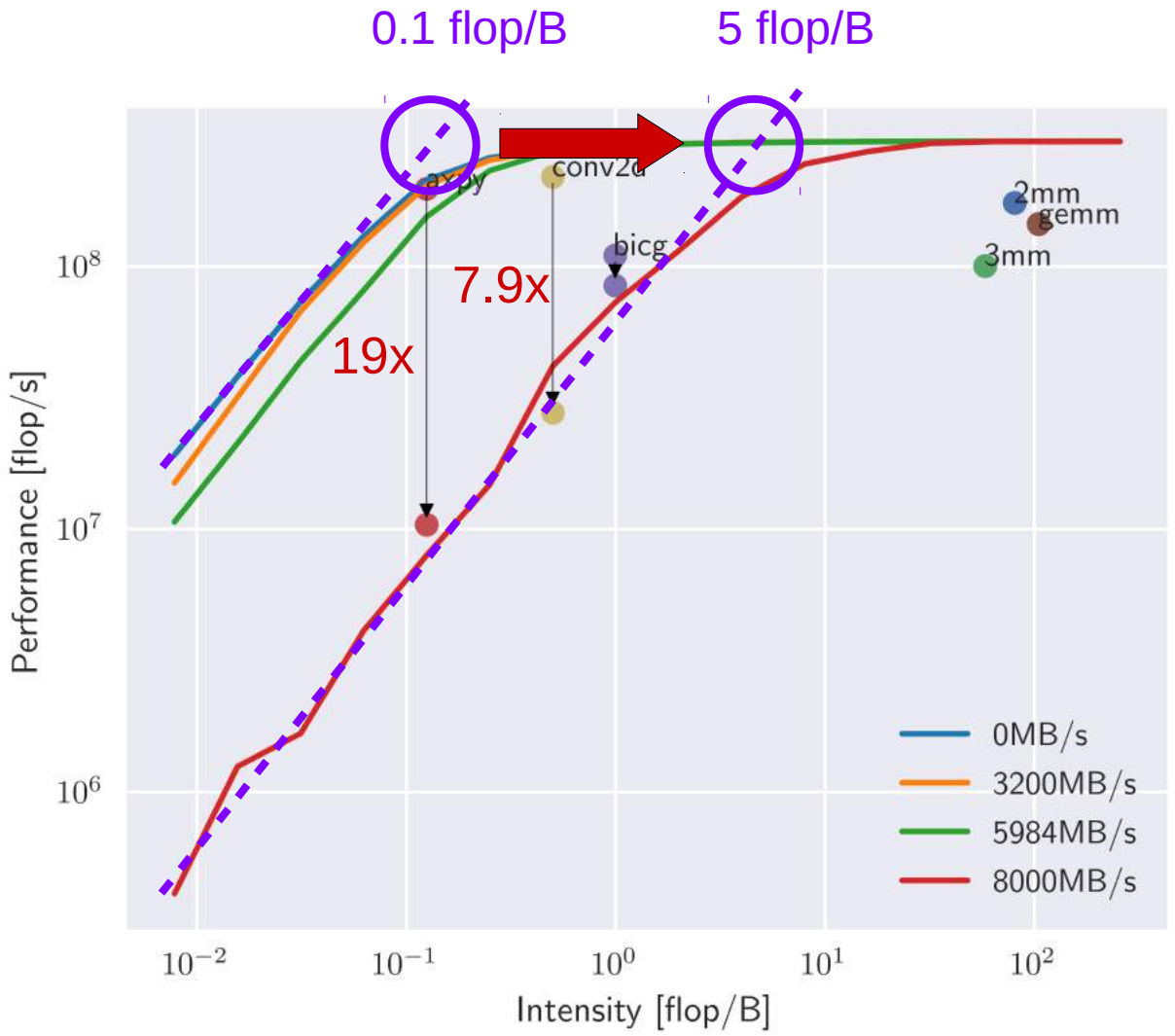
Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.27	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



Quantitative Impact

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

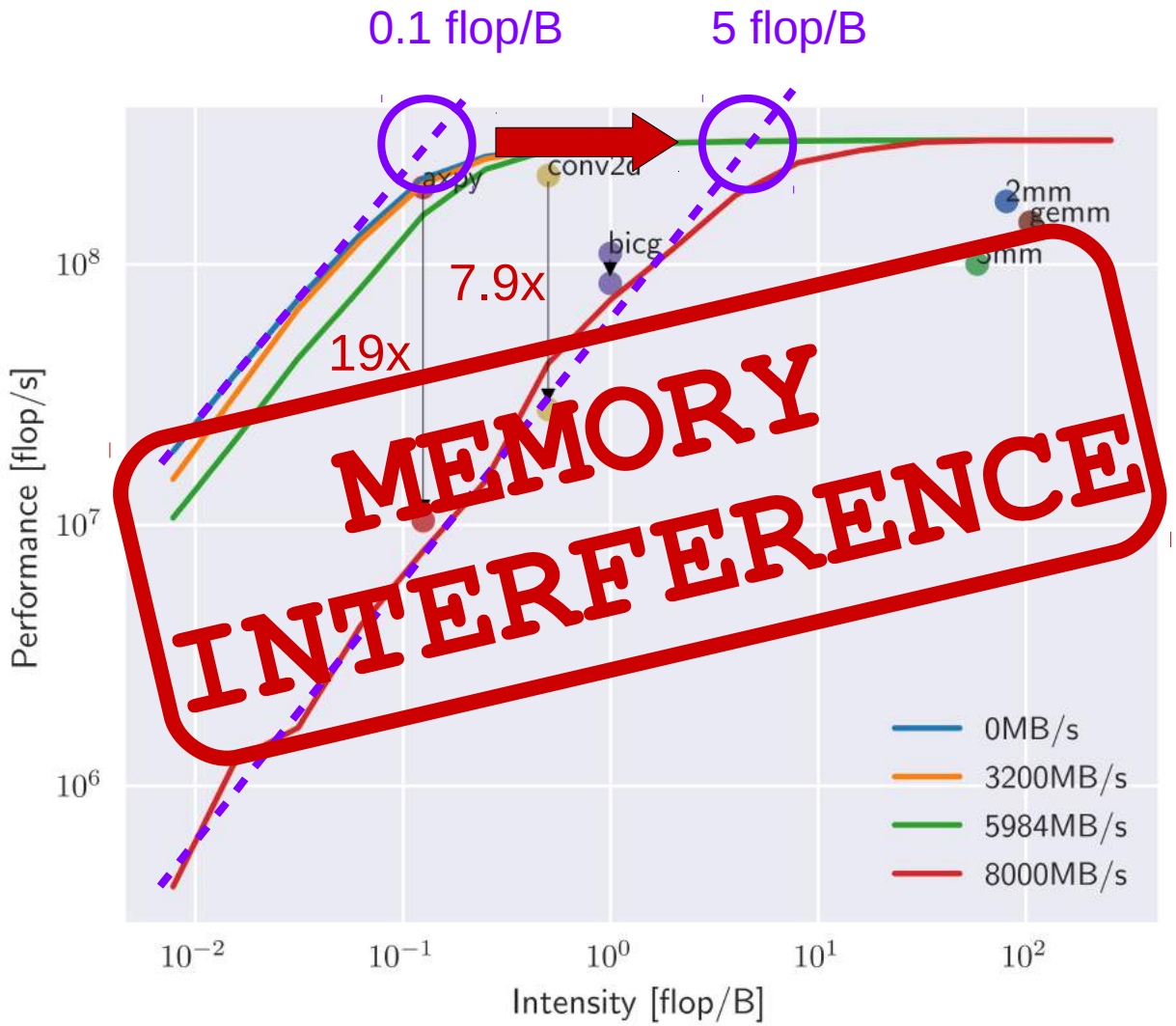
Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.27	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



Quantitative Impact

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.27	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

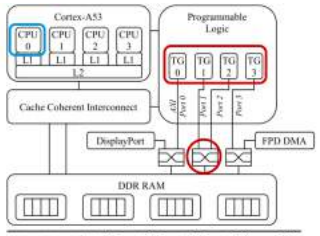
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H^{TG}_{01}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference

Algorithm 1: `stride` with intensity control.

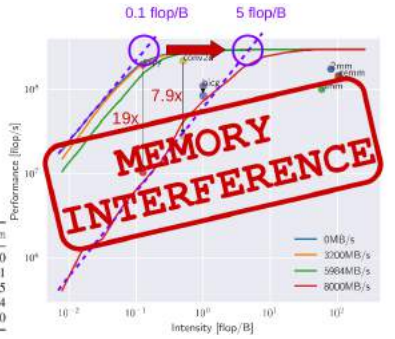
Data: vectors `X,Y` of length `n`, and a scalar `k`

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3 |   for j=0;j<k;j++ do
4 |   |   Y[i]+=X[i];
5 |   end
6 end
```

QUANTITATIVE IMPACT

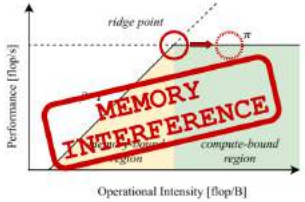
- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.22	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



NEW INSIGHTS

- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, 2009.

[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

SUMMARY AND CONCLUSION

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x
  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

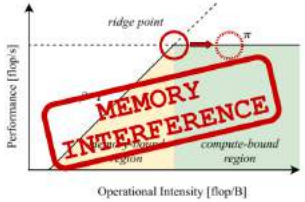
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

NEW INSIGHTS

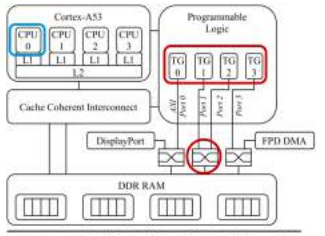
- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H^{TG}_{01}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference

Algorithm 1: `stride` with intensity control.

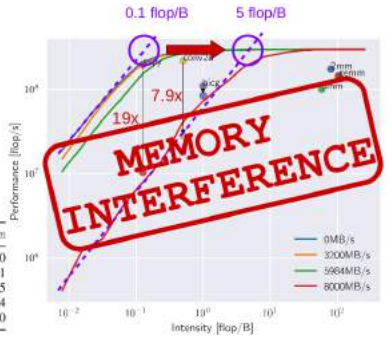
Data: vectors `X,Y` of length `n`, and a scalar `k`

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3     for j=0;j<k;j++ do
4         Y[i]+=X[i];
5     end
6 end
```

QUANTITATIVE IMPACT

- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.22	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



SUMMARY AND CONCLUSION

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x
  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, 2009.

[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

Summary and Conclusion

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results

  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation

  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x

  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

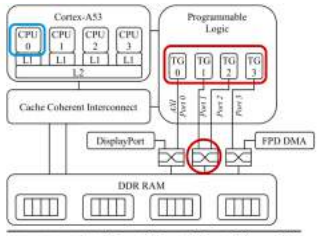
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2: Shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
  - Benchmarks with varying intensities
  - Memory-bound synthetic benchmark



Injected BW	H^{TG}_{01}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

- Synthetic benchmark: `stride`
  - Find configuration of worst performance
  - Cache misses every `k`-th memory access
  - Intensity control to measure rooflines under growing interference

Algorithm 1: `stride` with intensity control.

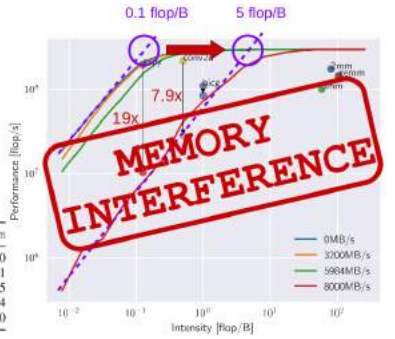
Data: vectors `X,Y` of length `n`, and a scalar `k`

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3 |   for j=0;j<k;j++ do
4 |   |   Y[i]+=X[i];
5 |   end
6 end
```

QUANTITATIVE IMPACT

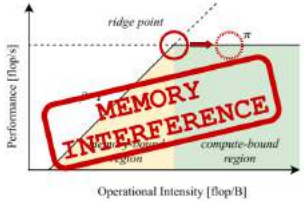
- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	1.02	1.00	0.98	1.05
5984 MB/s	1.00	1.00	1.22	1.01	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	7.91	1.00



NEW INSIGHTS

- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, 2009.

[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

SUMMARY AND CONCLUSION

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter growth from 1.2 to 10x
  3) Track ridge point behaviour
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

Analyzing Memory Interference of FPGA Accelerators on Multicore Hosts in Heterogeneous Reconfigurable SoCs

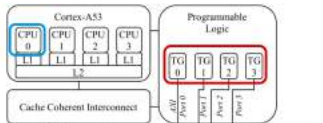
Maxim Mattheeuws, Björn Forsberg, Andreas Kurth, Luca Benini

BACKGROUND & MOTIVATION

Reprogrammable heterogeneous SoCs with high compute power are attractive for embedded applications, but all face a similar problem

DESCRIPTION

- Xilinx ZCU102
  - FPGA Traffic generators (TG)
  - Enabled/disabled individually
  - TG1 and TG2 shared port
  - Up to 8GB/s traffic to DRAM
- Cortex-A53 CPU
  - 32KiB data and instruction L1
  - 2MiB L2 cache
- Synthetic benchmarks with varying intensities
- Memory-bound synthetic benchmark: stride
  - Configuration for worst performance
  - Cache misses every k-th memory access
  - Intensity control to measure rooflines under growing interference



Injected BW	H^{TG}_{00}	H^{TG}_{12}	H^{TG}_{013}	H^{TG}_{012}	H^{TG}_{0123}
0 MB/s	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.17	1.12	1.11	1.12	1.12
3200 MB/s	1.26	1.29	1.33	1.21	1.31
5984 MB/s	2.51	3.01	2.09	2.03	1.62
8000 MB/s	4.37	6.51	11.44	6.22	25.9

Algorithm 1: `stride` with [illegible] control.

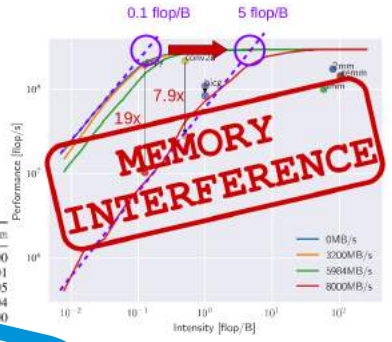
Data: vectors X,Y of length n, and a [illegible] k

```
1 stride s=16;
2 for i=0;i<n;i+=s do
3     for j=0;j<k;j++ do
4         Y[i]+=X[i];
5     end
6 end
```

QUANTITATIVE IMPACT

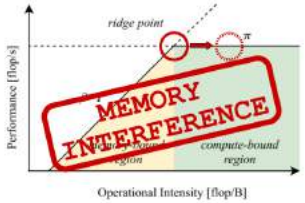
- Jitter:
  - Maximal deviation from median
  - 1.2x in non-interfered case
  - 10x with maximal interference
- Benchmark slowdown

Injected BW	2mm	3mm	axpy	bicg	conv2d	gemm
0 MB/s	1.00	1.00	1.00	1.00	1.00	1.00
320 MB/s	1.00	1.00	0.98	1.00	1.00	1.01
3200 MB/s	0.99	1.00	[illegible]	1.00	0.98	1.05
5984 MB/s	1.00	1.00	[illegible]	[illegible]	0.99	1.04
8000 MB/s	1.00	1.00	19.00	1.30	[illegible]	1.00



NEW INSIGHTS

- Analyze memory interference
  - State-of-the-Art Xilinx UltraScale+
- Up to 26x performance loss
  - 19x with real-world benchmark



- Model to characterize accelerator interference on CPU
  - Based on the roofline model[1]
  - Measurement-based extension with interference and worst case
  - Track ridge point

SUMMARY AND CONCLUSION

- Up to 19x performance loss of real-world benchmarks
- Novel degradation characterization methodology and results
  1) Find worst interference configuration using `stride` with minimal intensity
     - 26x Performance degradation
  2) Measure rooflines with increasing interference using `stride`
     - Jitter grows from 1.2 to 10x
  3) Track ridge point behavior
     - Increase from 0.1 to 5 flop/B
- Determine counter-measures such as *PREM*[3],[4] or *MemGuard*[5]
- Measurement based, as opposed to model-based[2]

References

[1] S. Williams *et al.*, "Roofline: an Insightful Visual Performance Model for Multi[illegible]rchitectures," *Commun. ACM*, 2009.

[2] S. Lee and C. Wu, "Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference," in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, "A Predictable Execution Model for COTS-Based Embedded Systems," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, "HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs," *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, "MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms," *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.

THANK YOU!

References

[1] S. Williams *et al.*, “Roofline: an Insightful Visual Performance Model for Multicore Architectures,” *Commun. ACM,* 2009.

[2] S. Lee and C. Wu, “Performance Characterization, Prediction, and Optimization for Heterogeneous Systems with Multi-Level Memory Interference,” in *IEEE Internat. Symp. on Workload Characterization*, 2017.

[3] R. Pellizzoni *et al.*, “A Predictable Execution Model for COTS-Based Embedded Systems,” *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2011.

[4] B. Forsberg, L. Benini, and A. Marongiu, “HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs,” *IEEE Transactions on Computers*, 2020.

[5] H. Yun *et al.*, “MemGuard: Memory Bandwidth Reservation System for Efficient Performance Isolation in Multi-Core Platforms,” *IEEE Real-Time and Embedded Technology and Applications Symposium*, 2013.