UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING


FAKULTÄT FÜR
ELEKTROTECHNIK UND
INFORMATIONSTECHNIK

# StreamGrid – an AXI-Stream-compliant Overlay Architecture

Christopher Blochwitz, León Philipp,
Mladen Berekovic, and Thilo Pionteck

ARC 2021
29 June – 1 July
Rennes, France

UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# Outline

- Motivation
- **Stream-Grid** – Overlay Architecture
- **Evaluation** – Impact of Design-Parameters
- **Case Study** – Database Queries
- Summary

# Motivation

- Application Specfic Hardware Accelerator
  - → Task Specific Hardware Accelerator
- Deviding an application in functional units (FU)
- Orchestration of FUs during run-time
- Partial reconfigurable partitions (RP)

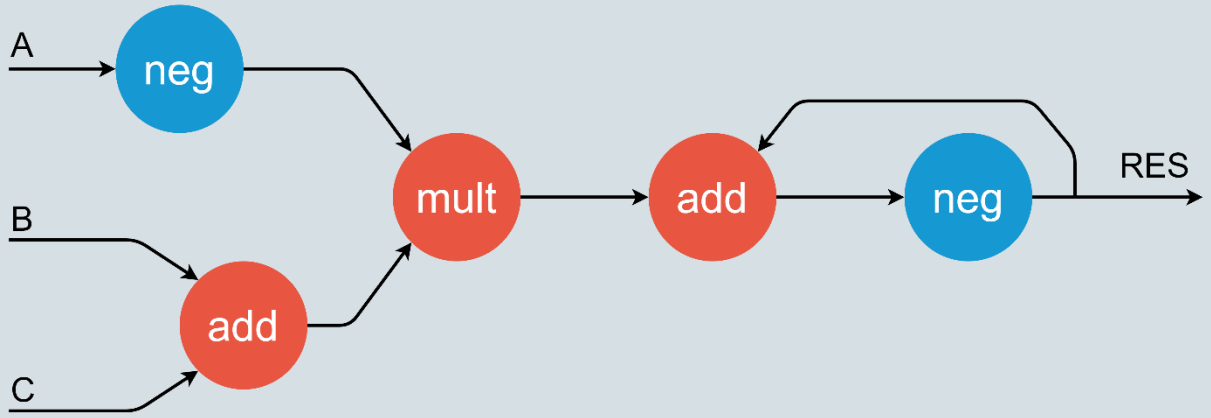
↳ **Overlay Architecture**

UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# Motivation

- Data-Flow-Graph
  - unary operator
  - binary operator
- Mapping to Overlay Architecture
- AXI-Stream compliant
  - wide set of IPs
  - Parameterizable Interface

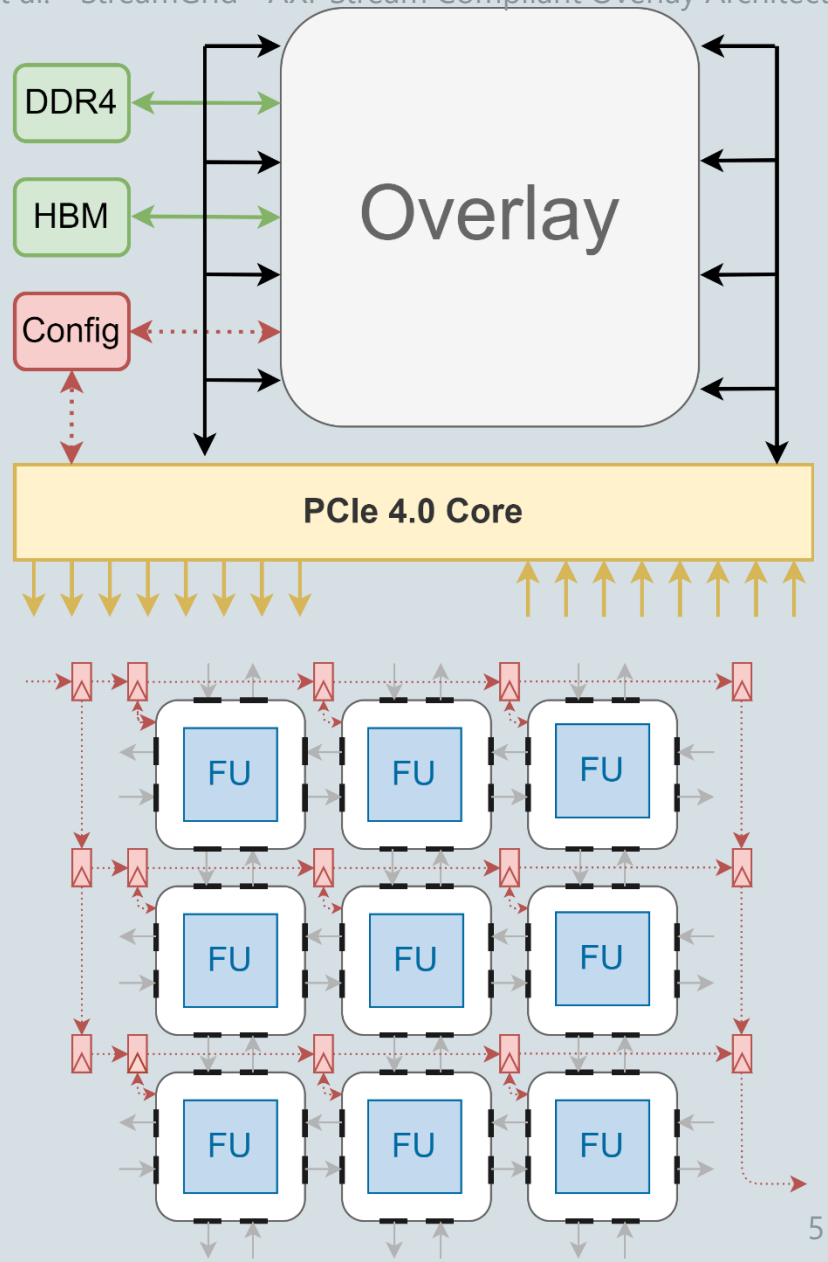
$$RES_i = -(RES_{i-1} + (-A * (B + C)))$$



UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# Stream-Grid – Overlay Architecture

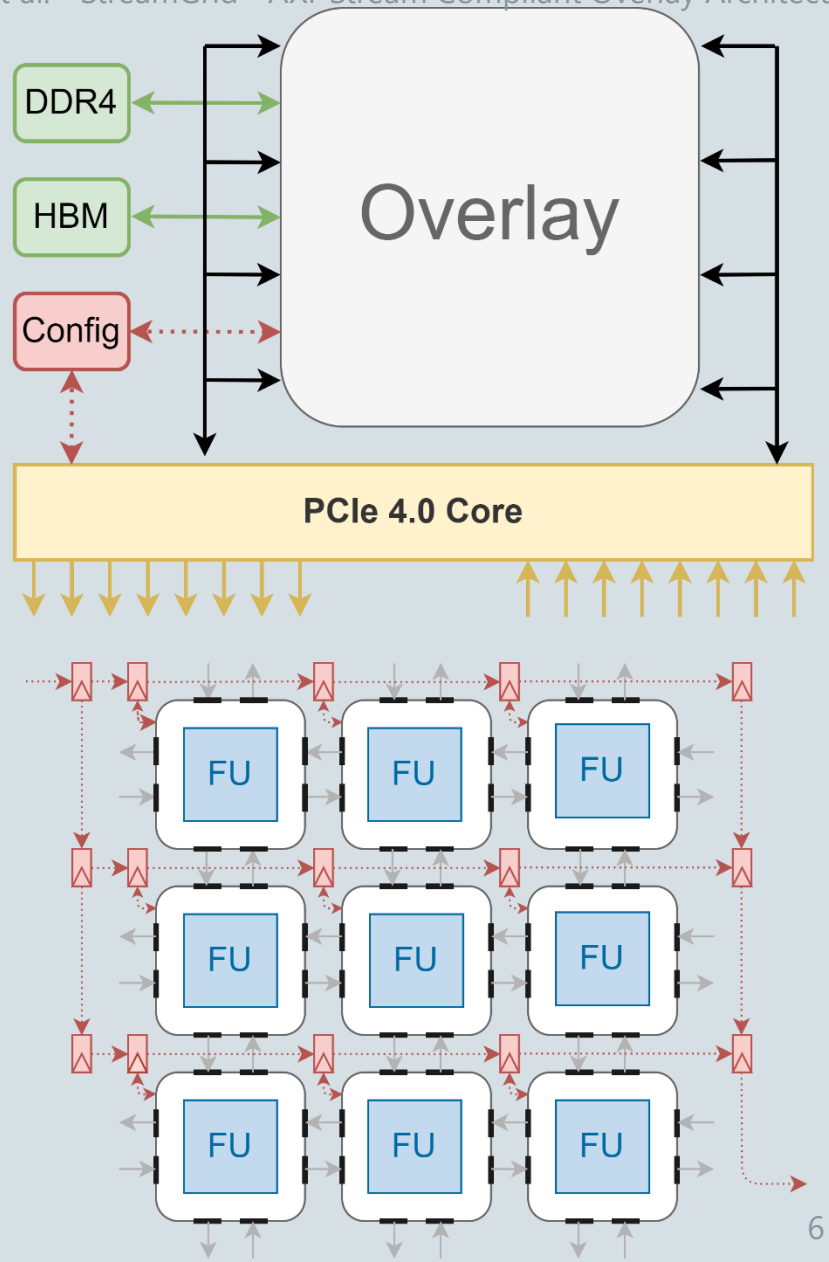
- Static Design
  - Memory Subsystem
  - PCIe 4.0 Interconnect
  - Configuration Network
- Parameter
  - Size of the Grid $n \times m$
  - AXI-Stream Interface
    - Data Width



UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# Stream-Grid – Overlay Architecture

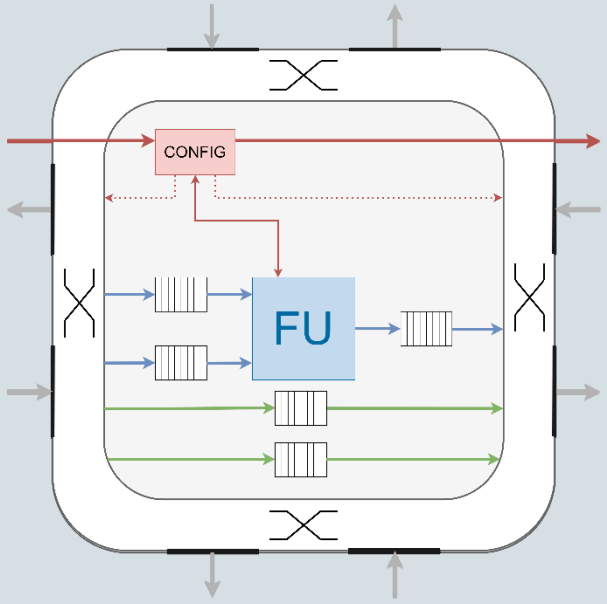
- Configuration & Monitoring
  - Packet based / y-x-Routing
  - Input/Output Switching
  - FU specific
    - Configuration
    - Monitoring

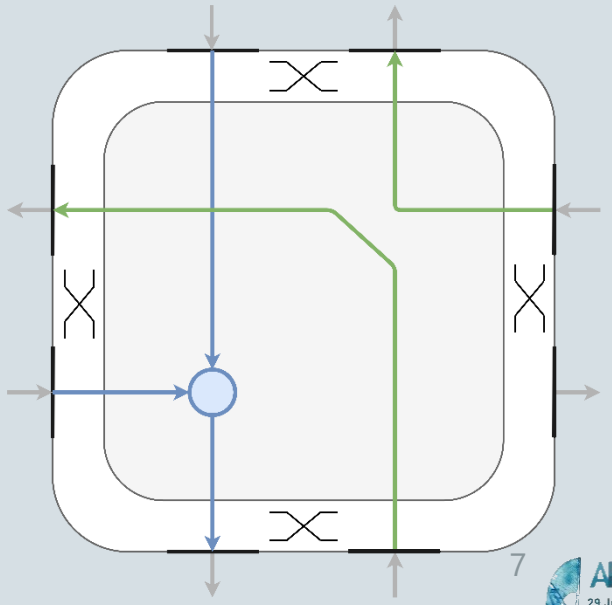


UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# Stream-Grid – Overlay Architecture

- Parameter
  - Buffer-Strategy – REG / FIFO BRAM / FIFO DistRAM
- Partial Reconfiguration
  - Utilization of the biggest FU → size of RP
  - Limitation (Xilinx): configuration frame
    - $width\ of\ 1\ element \times high\ of\ clock\ region$
- Advanced Routing mechanism
  - 1/2/3 parallel paths in one tile



# **Evaluation** – Impact of Design-Parameters

- System Configuration
  - Alveo U280
  - AMD EPYC 24-Core / 256 GB RAM
  - Ubuntu 20.04
- Default Parameters

| Parameter | default Value |
|---|---|
| Buffer-Strategy | REG |
| Grid-Size | 4 x 4 |
| Data Width | 256 bit |

UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# **Evaluation** – Impact of Design-Parameters

- Buffer-Strategy

default: 4x4 – 256 bit

| **Total** | **LUT** 1 303 680 | **Flip-Flop** 2 607 360 | **BRAM** 2 016 | **$F_{max}$** (MHz) |
|---|---|---|---|---|
| REG | 9 552 | 19 552 | 0 | 645 |
| FIFO dist. RAM | 32 078 | 53 858 | 0 | 655 |
| FIFO BRAM | 25 636 | 43 194 | 192 | 470 |

2.2 – 3.2 fewer Ressources
37% faster

- dist. RAM – 32 depth
- BRAM – 512 depth

**REG** or **BRAM** depending on application

UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# **Evaluation** – Impact of Design-Parameters

- Grid-Size

default: REG –256 bit

| Total | LUT 1 303 680 | per tile | Flip-Flop 2 607 360 | per tile | $F_{max}$ (MHz) |
|---|---|---|---|---|---|
| 2 x 2 | 2 125 | 531 | 4 053 | 1 013 | 683 |
| 3 x 3 | 5 097 | 566 | 10 354 | 1 150 | 635 |
| 4 x 4 | 9 552 | 597 | 19 552 | 1 222 | 645 |
| 5 x 5 | 15 420 | 617 | 31 694 | 1 268 | 640 |
| 6 x 6 | 22 733 | 631 | 46 758 | 1 299 | 640 |
| 7 x 7 | 31 433 | 641 | 64 748 | 1 321 | 684 |

Linear increas – Bundary effect

Constant high Frequency

2.4% / 2.5% Overall Ressources

**No limitiations** caused by StreamGrid

UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# **Evaluation** – Impact of Design-Parameters

- Data-Width

default: REG – 4x4

| Total | LUT 1 303 680 | per tile | Flip-Flop 2 607 360 | per tile | $F_{max}$ (MHz) |
|---|---|---|---|---|---|
| 32 | 2 131 | 133 | 4 848 | 303 | 752 |
| 64 | 3 194 | 200 | 6 997 | 437 | 687 |
| 128 | 5 327 | 333 | 11 165 | 698 | 697 |
| 256 | 9 552 | 597 | 19 552 | 1 222 | 645 |
| 512 | 18 612 | 1 163 | 40 618 | 2 538 | 630 |

Decreasing Frequency but still high

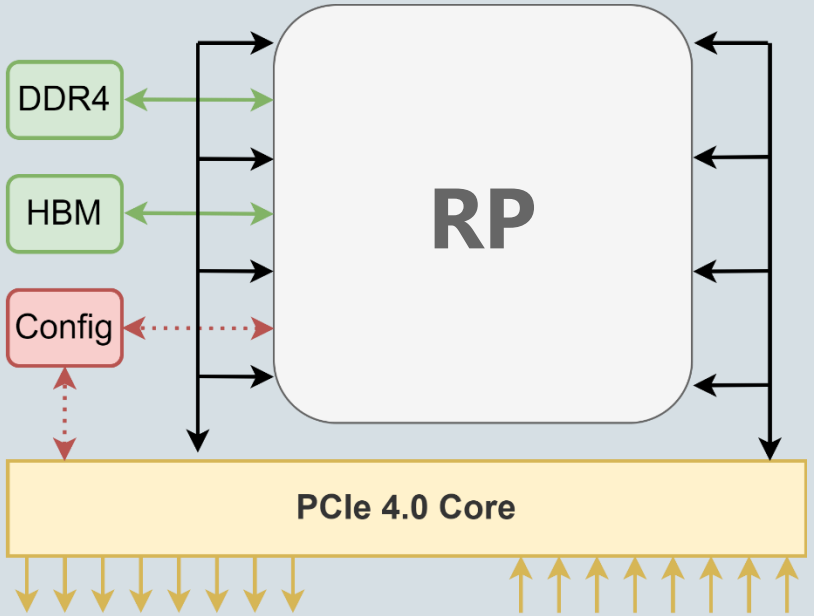
Data-Width depending on application

UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# **Case Study** – Database Queries

- Static Design by Werner et al. [11][13]
  - **One RP**
  - Procedure:
    - Query transformed to DFG
    - DFG to VHDL via makro
    - Starting Synthesis → long response time

↳
  - DFG operator to FU
  - Query to Configuration-Stream




[11] Stefan Werner: Hybrid Architecture for Hardware-accelerated Query Processing in Se mantic Web Databases based on Runtime Reconfigurable FPGAs. Ph.D. thesis, (2016)
[13] Werner, S., Heinrich, D., Pionteck, T., Groppe, S.: Semistatic operator graphs for accelerated query execution on FPGAs. Microprocessors and Microsystems (8 2017)

UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

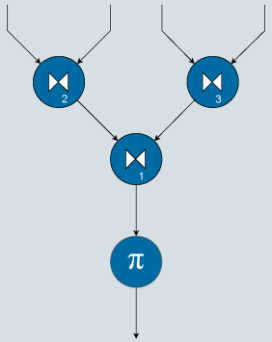
# Case Study – Database Queries

- SP²B Benchmark

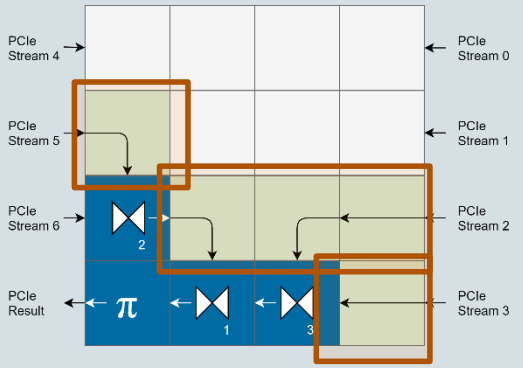
## Querry 3

```
1  #Get all articles with title, no. of pages and creator.
2  SELECT ?article ?title ?pages ?creator WHERE {
3    ?article rdf:type bench:Article .
4    ?article dc:title ?title .
5    ?article swrc:pages ?pages .
6    ?article dc:creator ?creator .
7  }
```

**DFG**



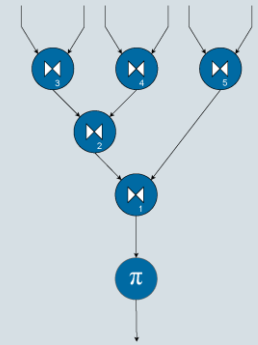
**Mapping**

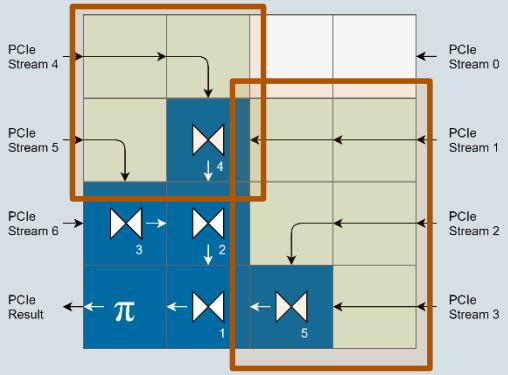



Multi-Stream: Tiles still usable for other queries

## Querry 5

```
1  #Get all articles with titles, no. of pages, creator,
       journal and pub. month.
2  SELECT ?article ?title ?pages ?creator ?journal ?
       month WHERE {
3    ?article rdf:type bench:Article .
4    ?article dc:title ?title .
5    ?article swrc:pages ?pages .
6    ?article dc:creator ?creator .
7    ?article swrc:journal ?journal .
8    ?article swrc:month ?month .
9  }
```





UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# **Case Study** – Database Queries

- Resource requirements - FU

| **Total** | **LUT 1 303 680** | **Flip-Flop 2 607 360** | **BRAM 2 016** | **$F_{max}$ (MHz)** |
|---|---|---|---|---|
| MergeJoin | 1 230 | 1 440 | 8 | 560 |
| Projection | 91 | 294 | 0 | 764 |
| RP | 2 880 | 5 760 | 8 | - - - |

Defines the RP

Defines the frequency of the overlay architecture

UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# **Case Study** – Database Queries

- Resource requirements - Queries

| | Total | LUT 1 303 680 | Flip-Flop 2 607 360 | BRAM 2 016 | $F_{max}$ (MHz) | Setup Time |
|---|---|---|---|---|---|---|
| Q3 | Static | 4 846 | 13 054 | 36 | 634 | 21:43 min |
| | Overlay | 16 841 | 32 608 | 24 | 560 | 1.79 ms |
| Q5 | Static | 7 699 | 19 104 | 60 | 586 | 30:29 min |
| | Overlay | 19 301 | 35 488 | 40 | 560 | 2.69 ms |

Higher utilization caused by the RPs

For Q3 – 13% slower clock frequency

StreamGrid **enables a suitable Setup Time**

UNIVERSITÄT ZU LÜBECK
INSTITUTE OF COMPUTER ENGINEERING

# **Case Study** – Database Queries

- Execution Time

| | Data Set | | Measured Execution Time | | | | Theoretical max Execution Time | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Static Design | | Overlay | | Static Design | | Overlay | |
| | [M] | [MB] | time [ms] | throu. [GB/s] | time [ms] | throu. [GB/s] | time [ms] | throu. [GB/s] | time [ms] | throu. [GB/s] |
| Q3 | 66 | 801 | 258 | 3.11 | 258 | 3.11 | 27.8 | 59.2 | 31.5 | 52.2 |
| | 131 | 1 578 | 501 | 3.15 | 501 | 3.15 | 47.6 | 68.2 | 53.8 | 60.2 |
| | 262 | 3 096 | 973 | 3.18 | 973 | 3.18 | 85.1 | 74.7 | 96.4 | 66.0 |
| Q5 | 66 | 838 | 268 | 3.12 | 268 | 3.12 | 26.7 | 64.5 | 27.8 | 61.8 |
| | 131 | 1 637 | 517 | 3.16 | 517 | 3.16 | 44.7 | 75.4 | 46.7 | 72.1 |
| | 262 | 3 198 | 1 006 | 3.18 | 1 006 | 3.18 | 79.3 | 82.7 | 83.1 | 79.0 |

Execution Time limited by PCIe

Execution Time slightly slower because of the lower clock frequency

# Summary

- StreamGrid – **AXI-Stream compliant** Overlay Architecture
- FU exchange on run-time by partial reconfiguration
- **Multi-Stream Architecture**
- **Parameterizable** – Buffer-Strategy / grid-size / data-width
- maximum Clock Frequency of **752 MHz**
- Utilization for 7 x 7 grid **<2.5%**

# Thank You!