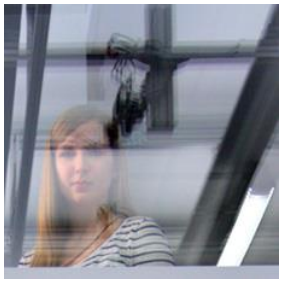
FPGA-Accelerated Heterogeneous Hyperscale Server Architecture for Next-Generation Compute Clusters

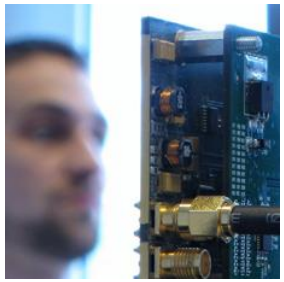
Rene Griessl, Peykanu Meysam, Jens Hagemeyer, Mario Porrmann
Bielefeld University, Germany

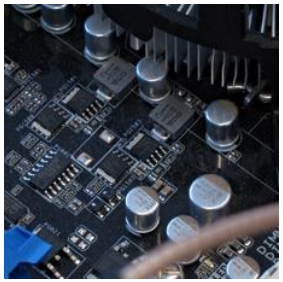
Stefan Krupop, Micha vor dem Berge
Christmann, Germany

Lars Kosmann, Patrick Knocke
OFFIS, Germany

Michał Kierzynka, Ariel Oleksiak
Poznan Supercomputing and Networking Center, Poland







FiPS – Motivation

Rising demand for computing power

- Complexity of calculations
- Increasing number of users

Energy consumption is rising

- 15 % of electrical energy consumption by computers
- CO_2 emissions

Systems have to be cooled

- Complexity/Cost rising

FiPS (Field Programmable Servers)

→ Founded by EC

→ Integrate FPGA / ARM

RECS|Box – Concept

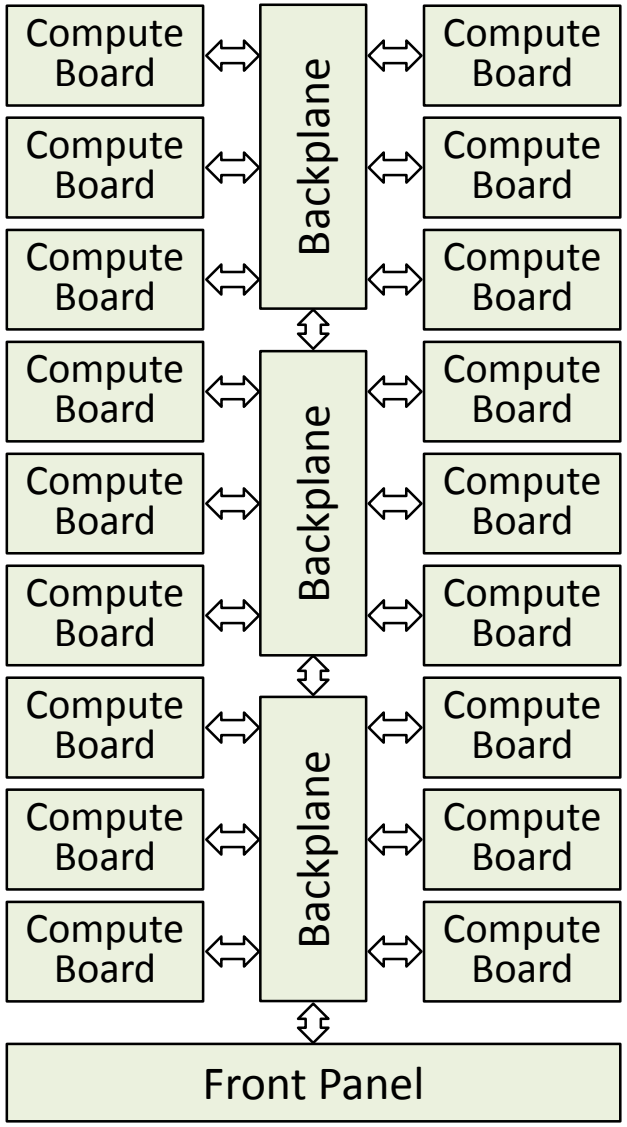
One single rack with up to 100 TFLOPS

- Up to 2,400 Microservers
- Up to 200 accelerators (e.g. GPGPU, Xeon Phi, FPGA)
- Up to 2.5 PB storage
- Up to 70 TB of RAM

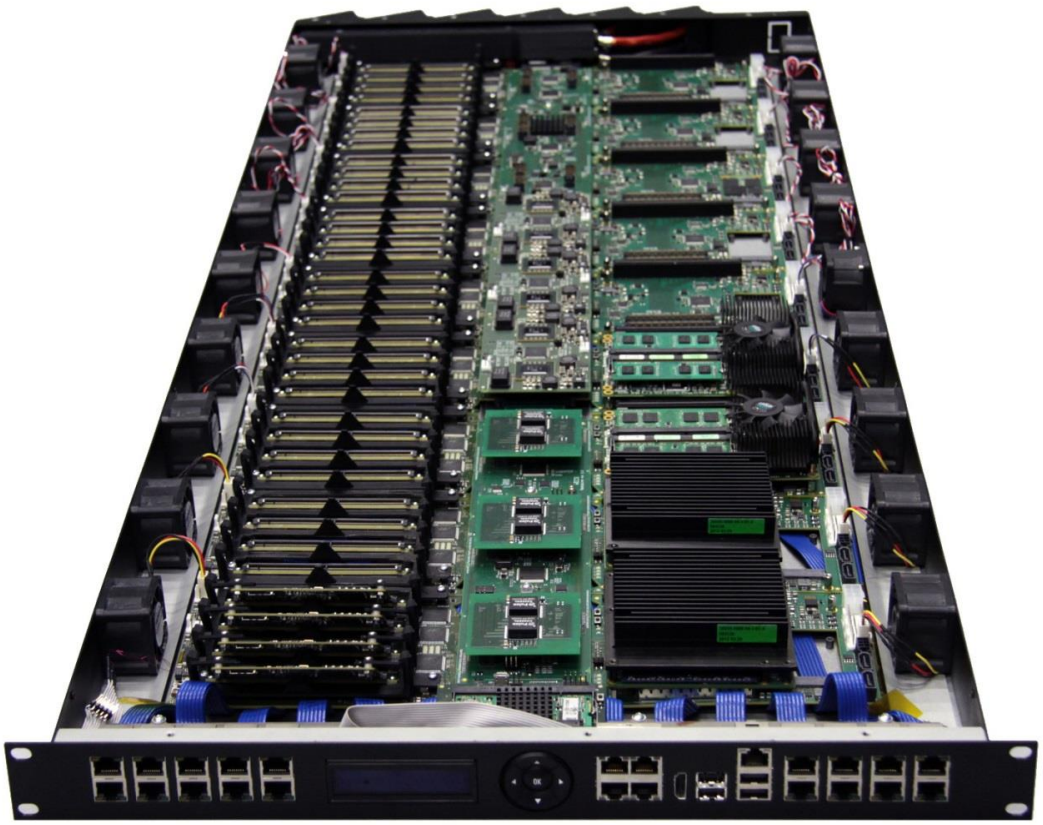
Main components

- **RCU: RECS | Box Compute Unit**
  - Contains CPU nodes, networking, management
- **RPU: RECS | Box Power Unit**
  - Intelligent power supply for RCUs
  - Can deliver power for multiple RCUs

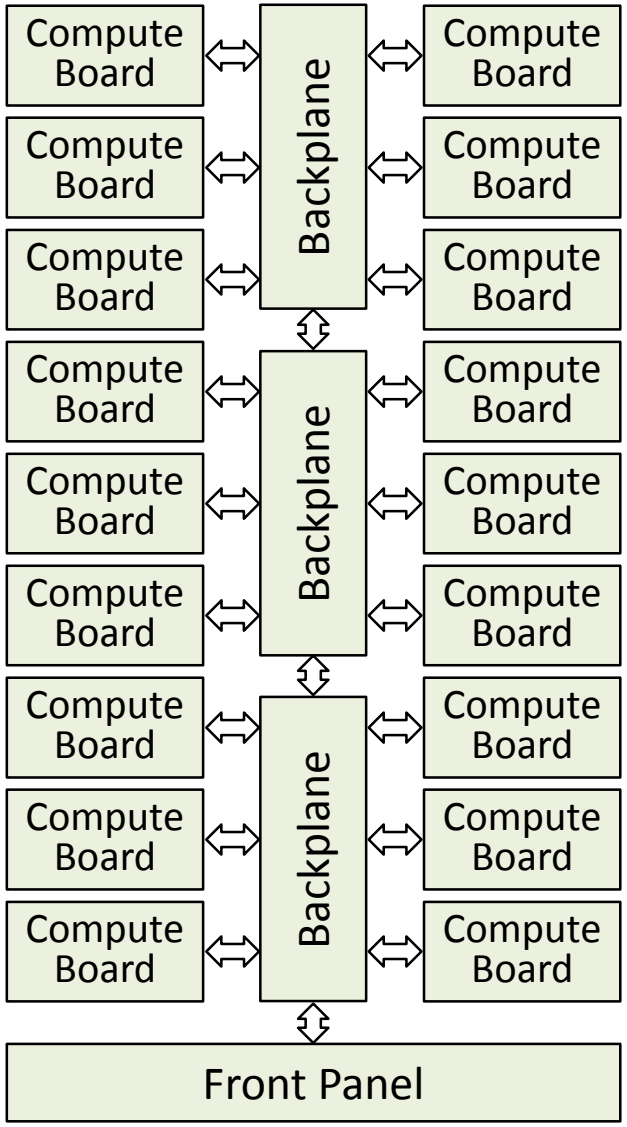
RECS|Box System Overview



- Modular Microserver Architecture
  - Arbitrary combinations of up to 72 Microservers in a single 1 RU Server



RECS|Box System Overview

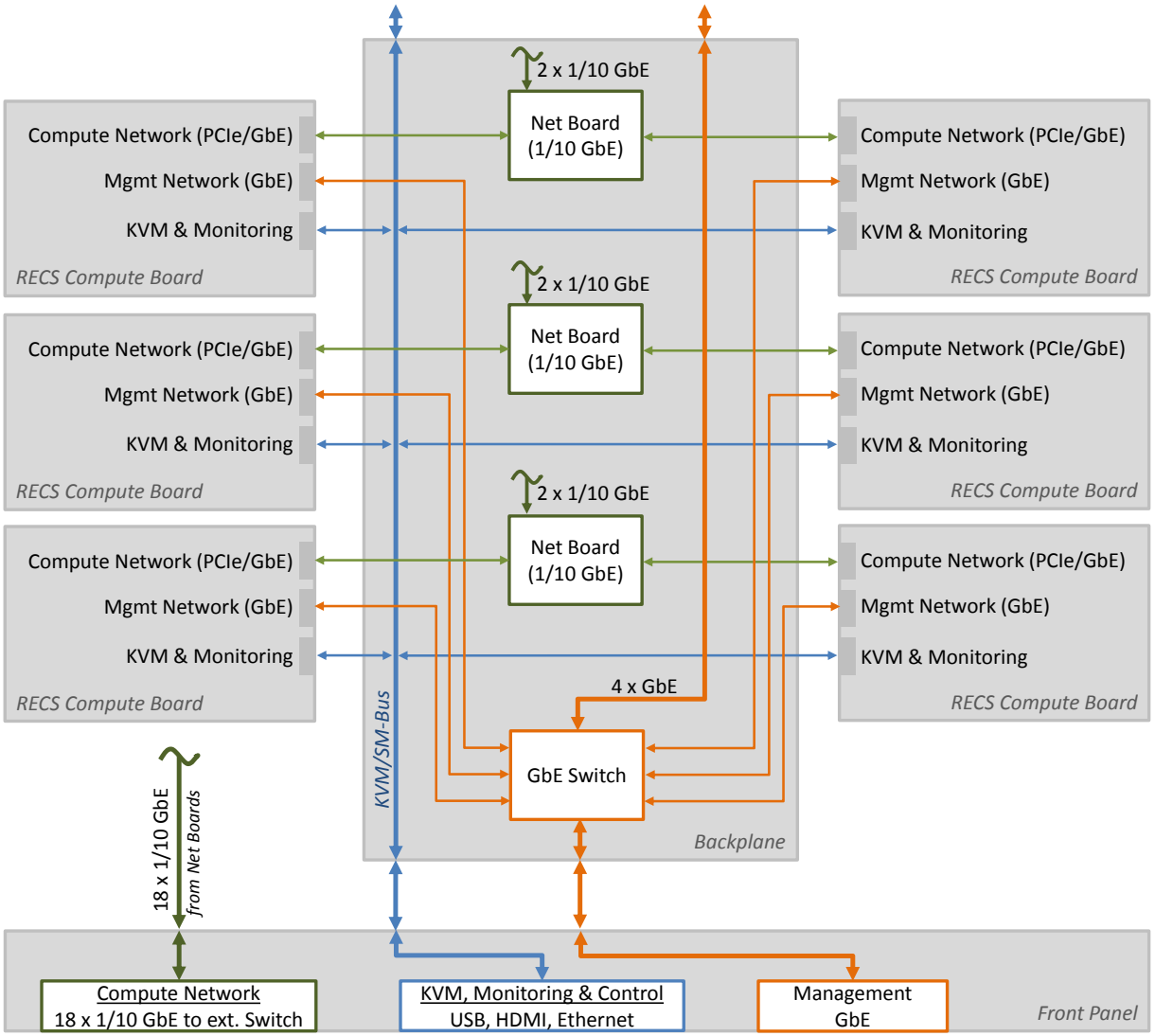


- Modular Microserver Architecture
  - Arbitrary combinations of up to 72 Microservers in a single 1 RU Server
  - Communication Backplane
    - Scalable from 6 to 18 compute boards
    - Flexible communication using dedicated Net-Boards
  - Compute Boards
    - Modular Architecture
    - Integrate microservers based on traditional CPUs and mobile CPUs
  - Hardware Accelerators
    - Integrated in specialized microserver
    - Attached via PCIe

RECS|Box Communication Backplane

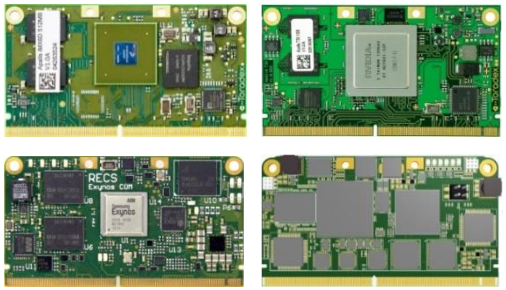
Three Levels of Interconnect

- Monitoring and Control
  - Distributed monitoring environment
  - Continuous sensing of volt., curr., temp., …
  - Integrated KVM
- Management
  - Gb Ethernet, switched on backplane
- High Throughput
  - 10 GbE
  - Infiniband



RECS|Box Microserver Overview

Apalis-based microservers

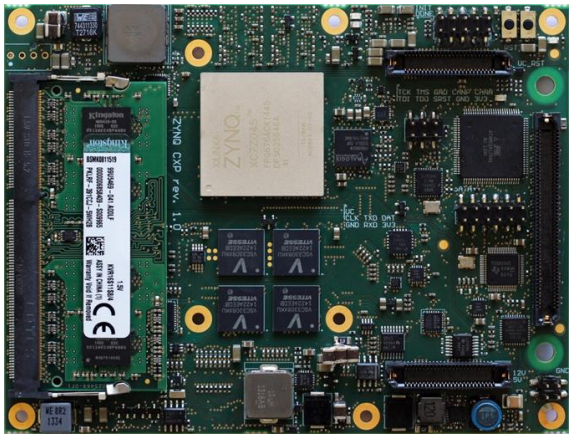


- ARM CPUs ranging from Cortex-A9 to Cortex-A15
- Xilinx Zynq-based module

COM Express-based microservers

- x86 modules range from Atom single cores to i7 quad core
- FPGA-based COM Express module developed, integrating Xilinx Zynq SoC

RECS|Box Microserver
Zynq COM Express Module



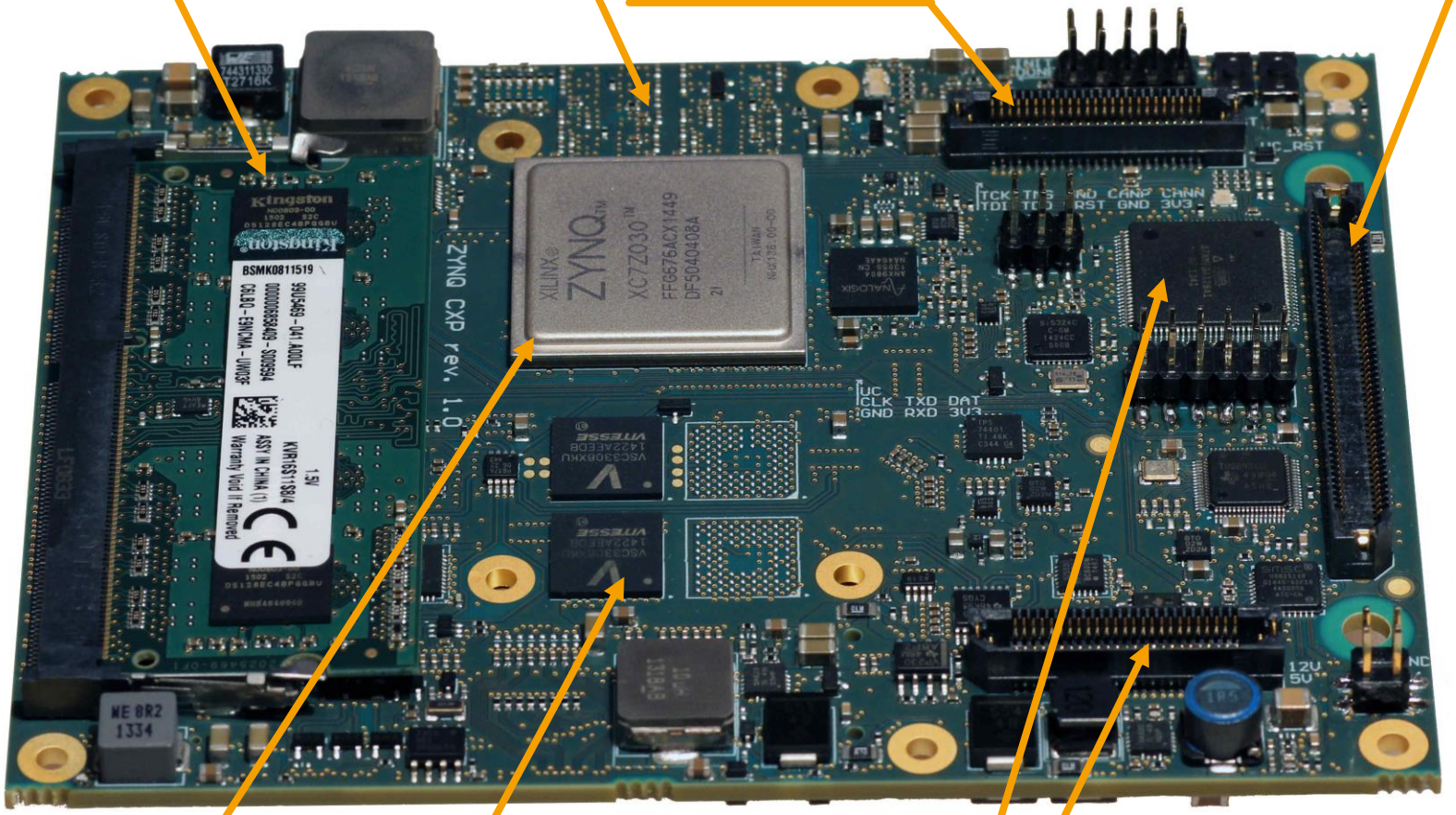
- Zynq-7000, ARM A9 Dual Core, 1 GHz
- Tightly integrated Programmable Logic
  - Used to extend Processing System
  - High performance ARM AXI interfaces
  - IP cores on PL
- Memory interfaces
  - 1 GByte of DDR3 (32-bit) PS Memory
  - up to 4 GByte DDR3 SO-DIMM module (64-bit-wide) PL Memory
  - eMMC memory (16 GByte)
  - Secure Digital (SD) card slot
- Management network: Gigabit Ethernet (PS)
- High-speed serial links (PCIe 2.1, 5 Gb/s)
- PCIe-based high toughput network can be implemented

Zynq-7000 Overview

	Zynq-7000 AP SoC Devices	Z-7010	Z-7015	Z-7020	Z-7030	Z-7035	Z-7045	Z-7100
Processing System	Processor Core	Dual ARM® Cortex™-A9 MPCore™						
	Processor Extensions	NEON™ & Single / Double Precision Floating Point						
	Max Frequency	866 MHz (-3) / 766 MHz (-2)			1 GHz (-3) / 800 MHz (-2)			
	Memory	L1 Cache 32KB I / D, L2 Cache 512KB, on-chip Memory 256KB						
	External Memory Support	DDR3, DDR2, LPDDR2, 2x QSPI, NAND, NOR						
	Peripherals	2x USB 2.0 (OTG), 2x Tri-mode Gigabit Ethernet, 2x SD/SDIO, 2x UART, 2x CAN 2.0B, 2x I2C, 2x SPI, 4x 32b GPIO						
Programmable Logic	Approximate ASIC Gates	~430K (30k LC)	~1.1K (74k LC)	~1.3M (85k LC)	~1.9M (125k LC)	~4.1M (275k LC)	~5.2M (350k LC)	~6.6M (444k LC)
	Block RAM	240KB	380KB	560KB	1,060KB	2,000KB	2,180KB	3,020KB
	Peak DSP Performance (Symmetric FIR)	100 GMACS	160 GMACS	276 GMACS	593 GMACS	1,334 GMACS	1334 GMACS	2622 GMACS
	PCI Express® (Root Complex or Endpoint)	-			Gen2 x4	Gen2 x8	Gen2 x8	Gen2 x8
	Agile Mixed Signal (XADC)	2x 12bit 1Msps A/D Converter						
I/O	Processor System IO	130						
	Multi Standards 3.3V IO	100	150	200	100	100	100	250
	Multi Standards High Performance 1.8V IO	-	-	-	150	150	150	150
	Multi Gigabit Transceivers	-	4 (6.25 Gbit/s)	-	4 (12.5 Gbit/s)	8 (12.5 Gbit/s)	8 (12.5 Gbit/s)	16 (12.5 Gbit/s)

[Xilinx, Zynq-7000 All Programmable SoC Overview, 2014]

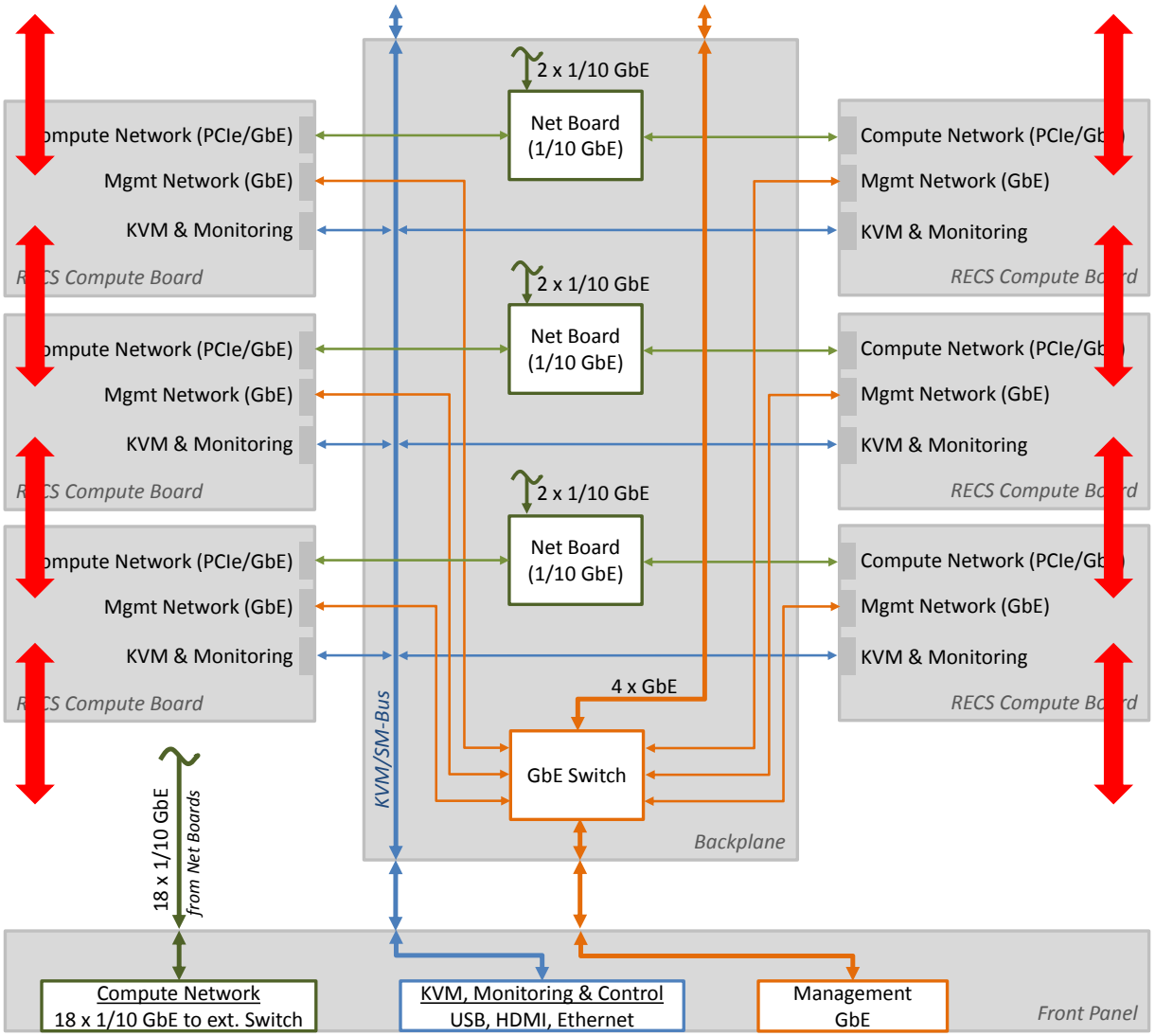
Zynq-7000 COM Express
Mechanical Overview



RECS|BOX Communication Backplane

Fourth Level of Interconnect

- Direct low latency
  - 8x 12.5 Gb/s
  - 300 ns
- High Throughput
  - 10 GbE
  - Infiniband
- Management
  - Gb Ethernet
- Monitoring and Control

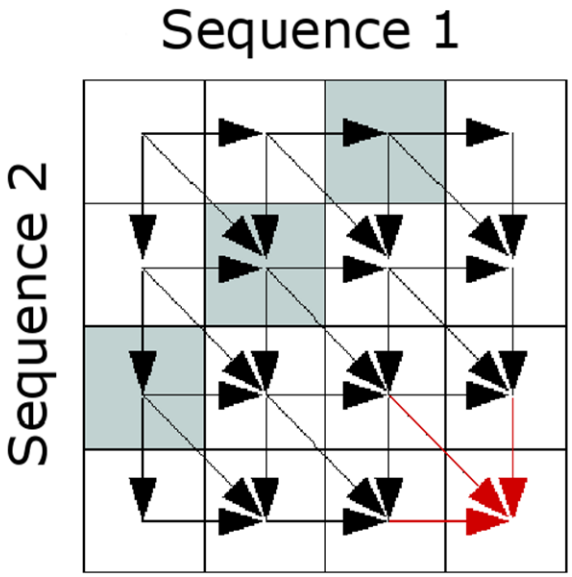


Benchmark DNA sequencing
GPU implementation

Needleman-Wunsch and dynamic programming (DP):

- Data dependencies: left, upper and diagonal elements are needed

$$H[i,j] = \max \begin{cases} H[i-1,j] - G_{penalty} \\ H[i,j-1] - G_{penalty} \\ H[i-1,j-1] + SM(s_1[i], s_2[j]) \end{cases}$$

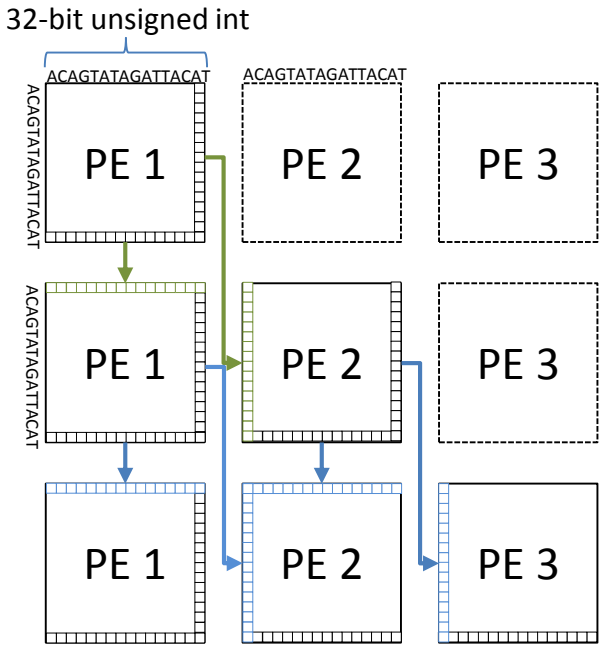


GPU implementation:

- The whole matrix is processed by a single GPU thread, thousands of threads work in parallel
- *MxN* matrix is divided into sub-matricies of *KxK* (*K* is the unroll factor)
- Up to 256 cells computed from a single data fetch
- Highly optimised for NVIDIA Fermi architecture using CUDA

Benchmark DNA sequencing
FPGA implementation

- Based on Vivado HLS
- Starting from basic C Needleman-Wunsch implementation
- Each PE calculates one sub-matrix of *KxK nucleoids*
- Systolic array style pipeline structure
- 11 PEs fit in Zynq-7045
- Zynq PS initializes and manages data flow
- Direct low latency links can be used for multi FPGA architecture



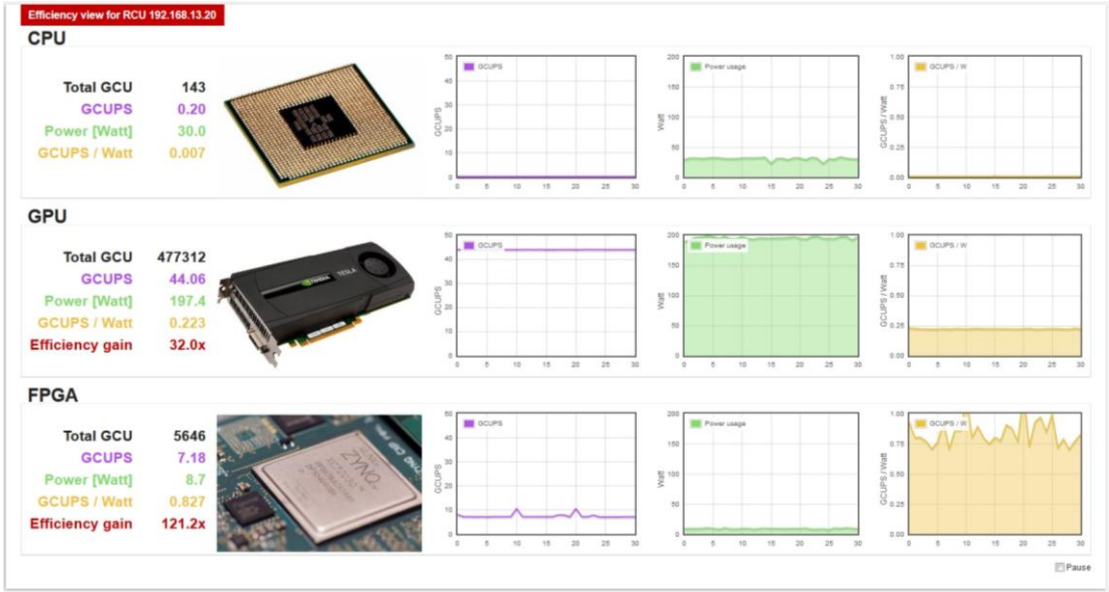
Total FPGA resources (11 PE)

	Used	Available	Utilization
Slice LUTs	209641	218600	95.90 %
LUT as Logic	208805	218600	95.51 %
LUT as Memory	836	70400	1.18 %
Slice Registers	171635	437200	39.25 %
Block RAM Tile	13	545	2.38 %

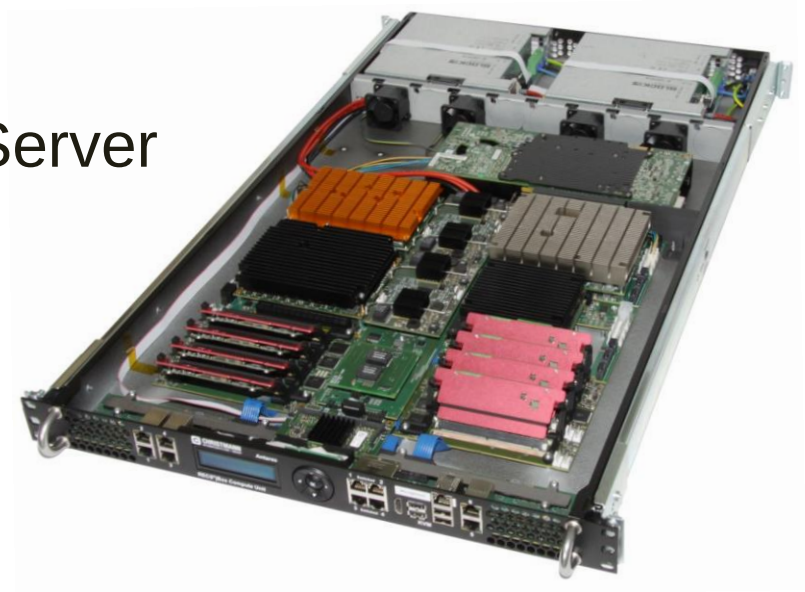
Benchmark DNA sequencing
Results

- Fastest implementation on Tesla GPU
- Most energy efficient implementation on FPGA
- Live Demo available
  - Visit Booth 2303

	GCUPS	Power [Watt]	GCUPS/ Watt	Efficiency gain
Intel i5-4400E	0.2	30	0.007	1
Tesla M2070	44.02	196.8	0.224	32
Tesla K40c	71.03	123.05	0.577	82.46
Zynq XC7Z045	7.08	8.7	0.814	116.26



Summary and Outlook



- RECS Resource Efficient Cluster Server
  - Integration of CPUs, embedded CPUs, GPGPUs and FPGAs
- Evaluation using DNA sequencing
  - HLS FPGA implementation provides maximum energy efficency

Outlook: EC – Horizon 2020 project

- M2DC – Modular Microserver Data Center
  - Architectural improvements focusing on communication
  - New Microservers (64-bit ARM, Zynq Ultrascale+, Hybrid Memory Cube)
  - Large scale testbeds / applications

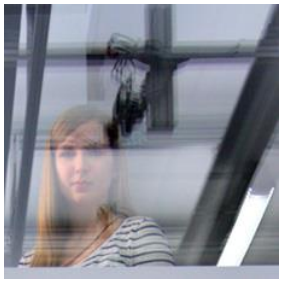
Universität Bielefeld

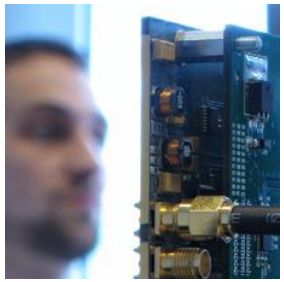
CITEC

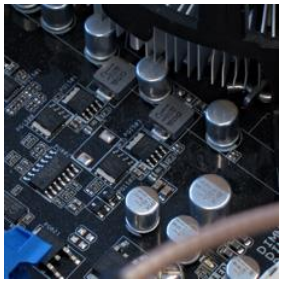
Many Thanks!

Visit RECS at booth 2303

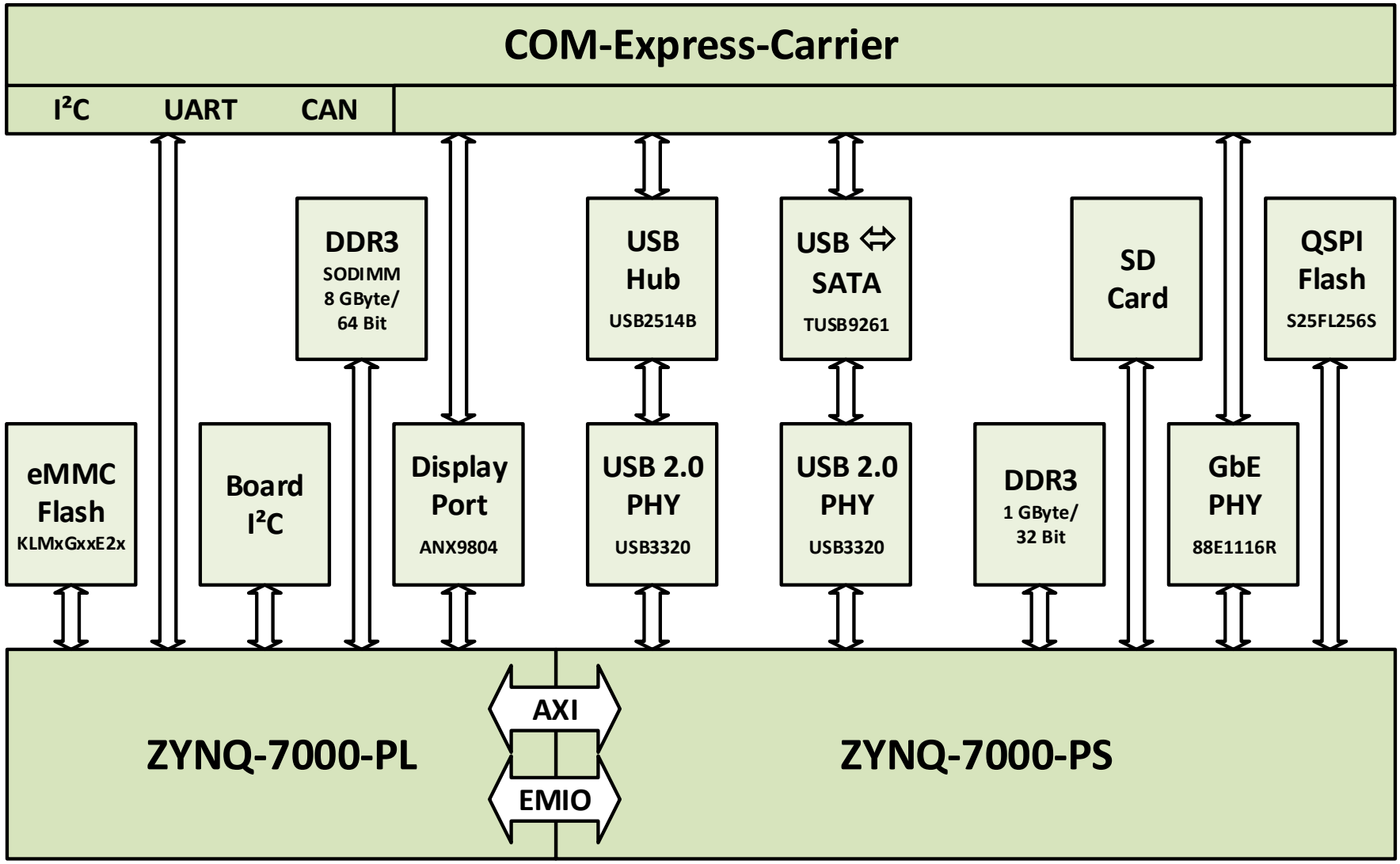
René Griessl
Cognitronics and Sensor Systems
Center of Excellence Cognitive Interaction Technology
Bielefeld University, Germany
rgriessl@cit-ec.uni-bielefeld.de
www.ks.cit-ec.uni-bielefeld.de







Zynq-7000 COM Express
System Architecture – COM Express Connector



→ Compatible with COM Express Type 6 Standards