NAAICE – Network-Attached Accelerators in Heterogenous Computing

S. Christgau[1] D. Everingham[1] T. Jaeuthe[2] M. de Lucia[3] M. Lübke[3,4] F. Mikolajczak[4]
D. Puhan[2] N. Schelten[5] B. Schnor[4] J. Spazier[2] B. Stabernack[4,5] F. Steinert[4,5]

[1]Zuse Institute Berlin

[2]Perfacct GmbH

[3]GFZ German Research Center for Geosciences Potsdam

[4]University of Potsdam

[5]Fraunhofer Heinrich-Hertz-Institute

PARS Workshop 2023, Aachen

Project Partners

- NAAICE is one of nine funded projects within the BMBF GreenHPC Call
- **Goal:** energy-efficient High Performance Computing
- two of them with ZIB participation (NAAICE and SeqAn@FPGA)
- NAAICE consortium:

University of Potsdam

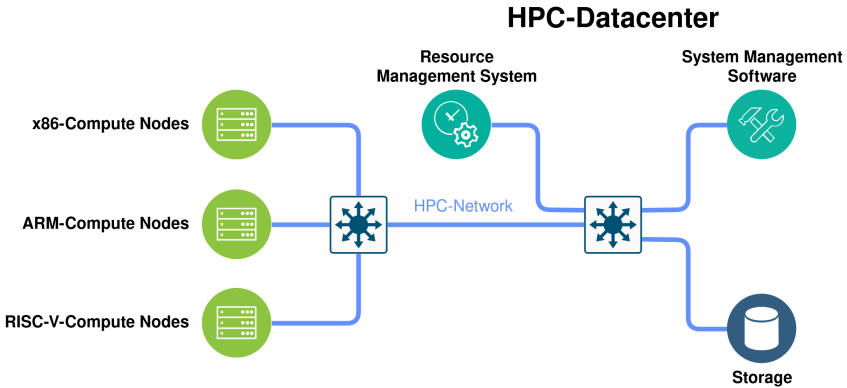
Fraunhofer Heinrich-Hertz-Institut, Berlin

Zuse-Institut Berlin

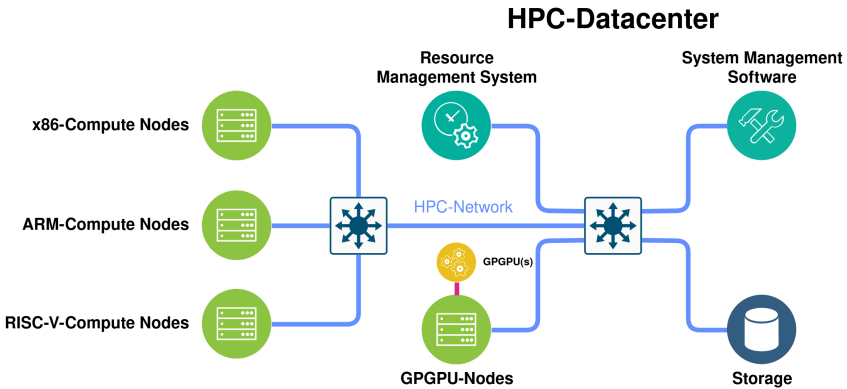
GeoForschungs-Zentrum Potsdam

PERFACCT GmbH, Potsdam

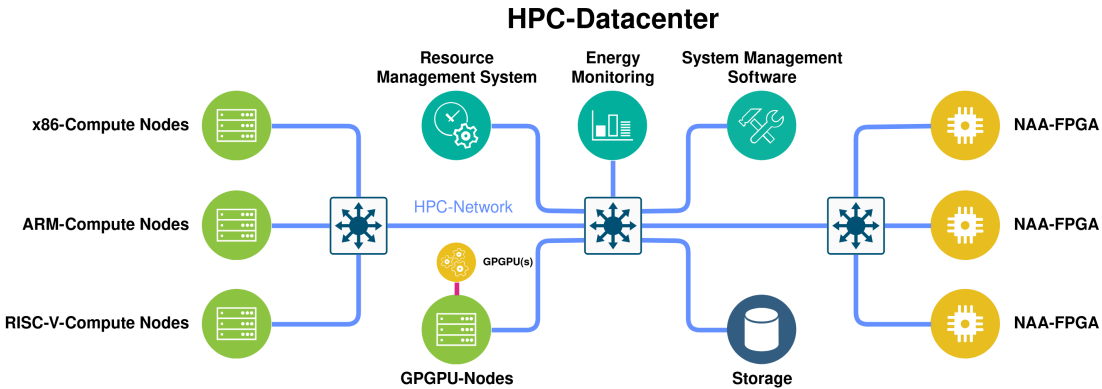
Motivation


HPC-Datacenter
Resource Management System
System Management Software
x86-Compute Nodes
ARM-Compute Nodes
RISC-V-Compute Nodes
HPC-Network
Storage

Motivation


HPC-Datacenter
Resource Management System
System Management Software
x86-Compute Nodes
ARM-Compute Nodes
RISC-V-Compute Nodes
HPC-Network
GPGPU(s)
GPGPU-Nodes
Storage

Motivation


HPC-Datacenter
Resource Management System
Energy Monitoring
System Management Software
x86-Compute Nodes
ARM-Compute Nodes
RISC-V-Compute Nodes
HPC-Network
GPGPU(s)
GPGPU-Nodes
Storage
NAA-FPGA
NAA-FPGA
NAA-FPGA

NAAICE approach for energy-efficient HPC

1. Network-Attached Accelerator (NAA)

Off-loading of compute-intensive tasks:

1. Integration of energy-efficient FPGA-based accelerators
2. NAA connected directly to the HPC network: 100 GBit/s RoCEv2 network (RDMA over Converged Ethernet) **avoiding the energy and resource overhead of a carrier system**

NAAICE approach for energy-efficient HPC

1. Network-Attached Accelerator (NAA)

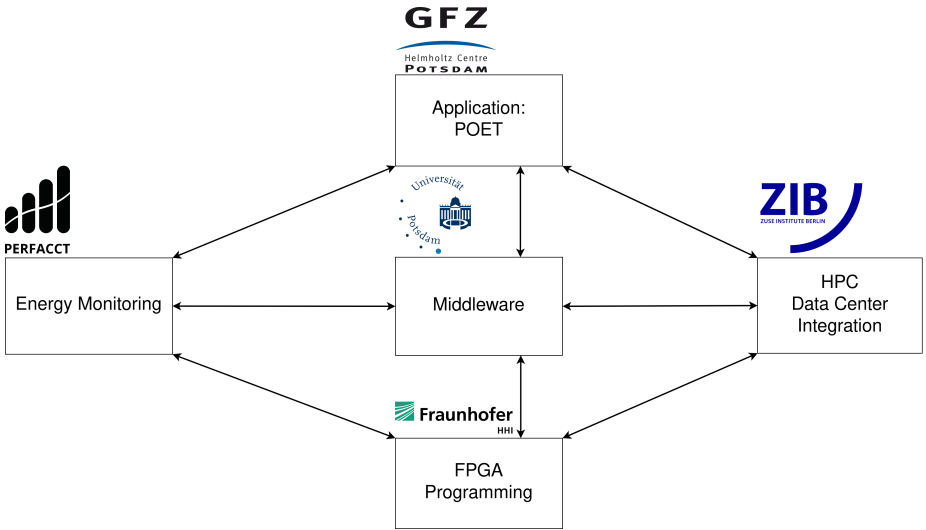
Off-loading of compute-intensive tasks:

1. Integration of energy-efficient FPGA-based accelerators
2. NAA connected directly to the HPC network: 100 GBit/s RoCEv2 network (RDMA over Converged Ethernet) **avoiding the energy and resource overhead of a carrier system**

2. Approximate Computing

Application-level: **KI-based surrogate model** instead of compute-intensive simulations

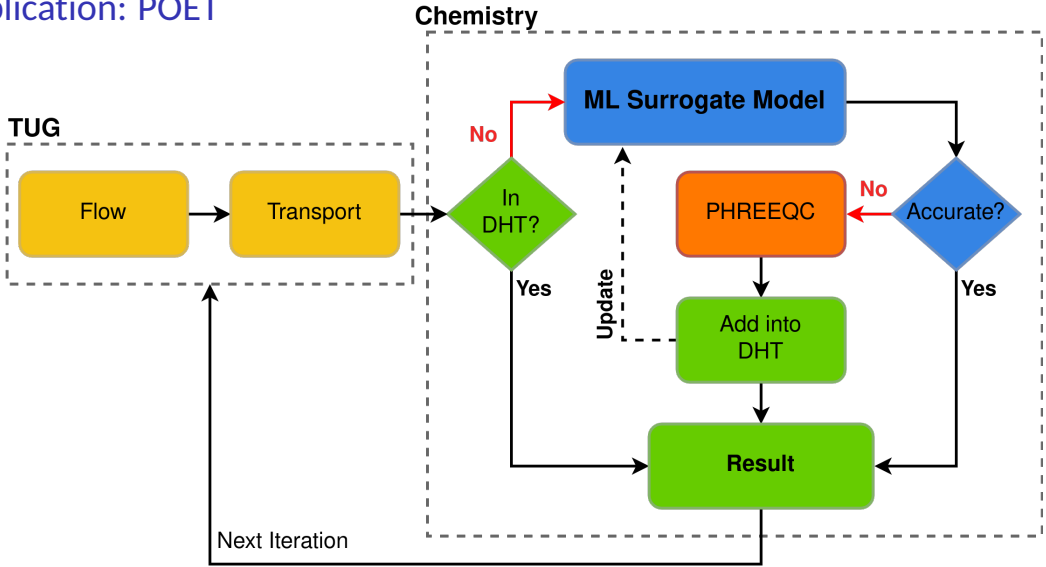
Responsibilities



Project Goals

1. Enabling Efficient Communication with NAAs in HPC environment
2. low usage barrier: integration into the *typical* HPC workflow like for example the integration of the NAA into the resource management system
3. energy-efficiency monitoring for HPC data center operators
4. validation of the NAAICE approach with the POET application

Application: POET

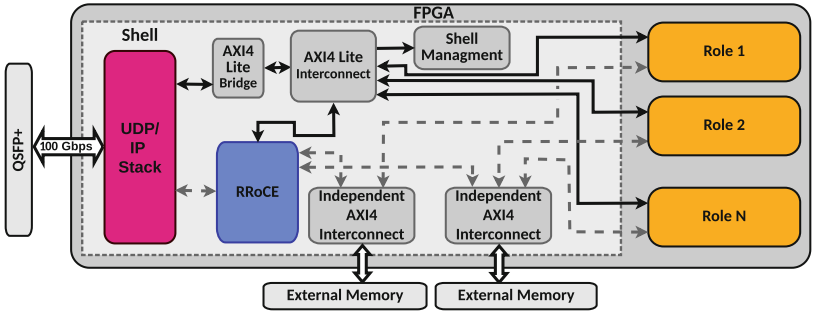


FPGA Framework

- Goal 1: Enabling Efficient Communication with NAAs in HPC environment

FPGA Framework

- Goal 1: Enabling Efficient Communication with NAAs in HPC environment
- Solution: RDMA-capable FPGA framework
  - Application-independent shell for enabling communication
  - Supported RDMA protocol: RoCEv2 = InfiniBand messages packed in UDP/IP/Ethernet frames
  - Offloaded computations in one of multiple accelerators roles/sockets (reconfigurable)



Communication with NAAs

- Communication Model: asynchronous remote procedure calls
  - Make use of FPGA framework's RDMA capabilities
  - Use case: long-running offloaded operations $\rightarrow$ asynchronous by design

Communication with NAAs

- Communication Model: asynchronous remote procedure calls
  - Make use of FPGA framework's RDMA capabilities
  - Use case: long-running offloaded operations $\rightarrow$ asynchronous by design

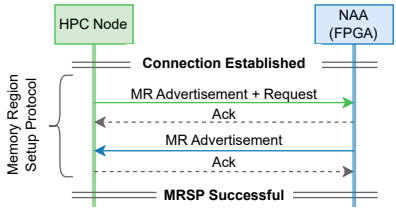
- Challenges:
  - Queue-pair management $\rightarrow$ managed by FPGA framework
  - Management of registered memory regions (MR)
  - Communication protocol for RPCs
  - Make it usable from application $\rightarrow$ API design

Solving Communication Challenges

- Memory Region Setup Protocol (MRSP)
  - Memory region = remotely accessible memory chunk
    $\rightarrow$ exchange of meta data required

Solving Communication Challenges

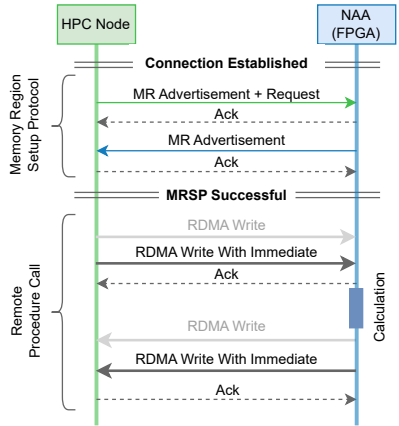
- Memory Region Setup Protocol (MRSP)
  - Memory region = remotely accessible memory chunk
    $\rightarrow$ exchange of meta data required
  - Advertisement of meta data from both sides
  - Symmetric memory regions between host and NAA
    $\rightarrow$ allows for exchange of inputs and outputs



Solving Communication Challenges

- Memory Region Setup Protocol (MRSP)
  - Memory region = remotely accessible memory chunk
    $\rightarrow$ exchange of meta data required
  - Advertisement of meta data from both sides
  - Symmetric memory regions between host and NAA
    $\rightarrow$ allows for exchange of inputs and outputs

- Performing the RPC
  - Transfer parameters via RDMA write (“put”)
  - Start computation with RDMA write with immediate
    $\rightarrow$ no further coordination needed
  - Result + completion notification via RDMA too

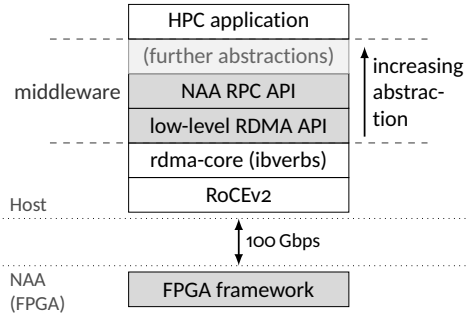


Application Programming Interface

- Goal 2: low usage barrier

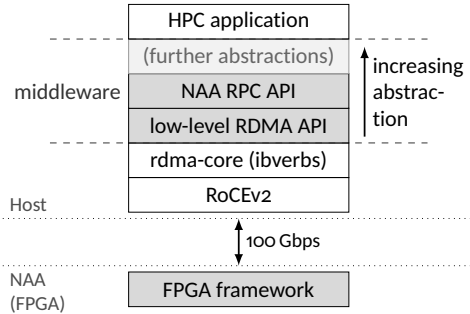
Application Programming Interface

- Goal 2: low usage barrier
- Middleware on top ibverbs/Linux RDMA stack
- handle-based, asynchronous design



Application Programming Interface

- Goal 2: low usage barrier
- Middleware on top ibverbs/Linux RDMA stack
- handle-based, asynchronous design




```
double *a = ..., *b = ..., *c = ...;
naa_param_t params[] =
    {{a, N * sizeof(*a)},
     {b, N * sizeof(*b)},
     {c, N * sizeof(*c)}};

// Instantiate an NAA connection.
naa_handle naa;
naa_create(FNCODE, &params, 3, &naa);

// Invoke the NAA routine.
naa_param_t in_param[] = {params[0], params[1]};
naa_param_t out_param[] = {param[2]};
naa_invoke(&in_params, 2, &out_params, 1, &naa);

int flag = 0;
while (!flag) {
    naa_test(&naa, &flag, ...);
    // Do other work while waiting on the NAA
}
```

Performance Evaluation

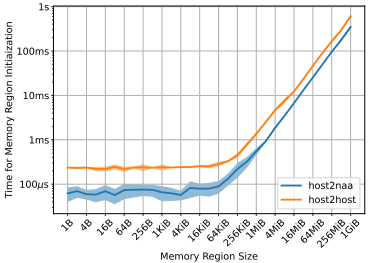
Evaluation of RDMA communication with NAA (FPGA framework)

- Time for Memory Region Setup Protocol, latency, bandwidth; Switched 100 Gbps connection
- Baseline: host-to-host (software prototype of FPGA framework)

Performance Evaluation

Evaluation of RDMA communication with NAA (FPGA framework)

- Time for Memory Region Setup Protocol, latency, bandwidth; Switched 100 Gbps connection
- Baseline: host-to-host (software prototype of FPGA framework)

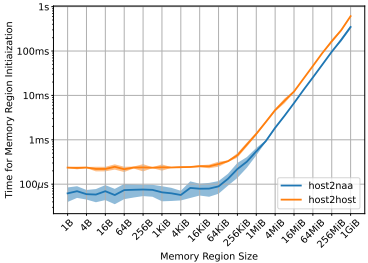


(a) Time for MRSP

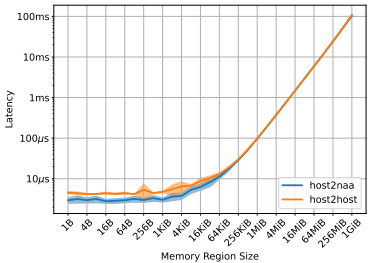
Performance Evaluation

Evaluation of RDMA communication with NAA (FPGA framework)

- Time for Memory Region Setup Protocol, latency, bandwidth; Switched 100 Gbps connection
- Baseline: host-to-host (software prototype of FPGA framework)



(a) Time for MRSP

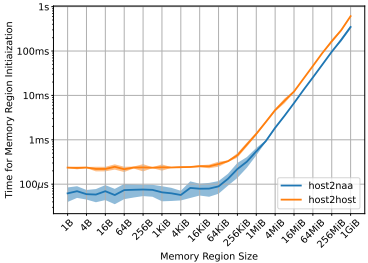


(b) Communication Latency

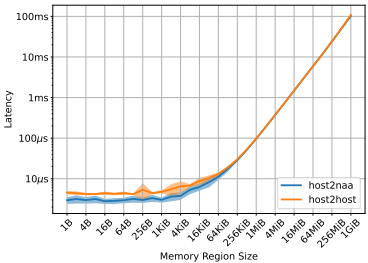
Performance Evaluation

Evaluation of RDMA communication with NAA (FPGA framework)

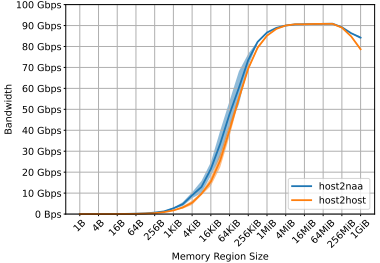
- Time for Memory Region Setup Protocol, latency, bandwidth; Switched 100 Gbps connection
- Baseline: host-to-host (software prototype of FPGA framework)



(a) Time for MRSP



(b) Communication Latency

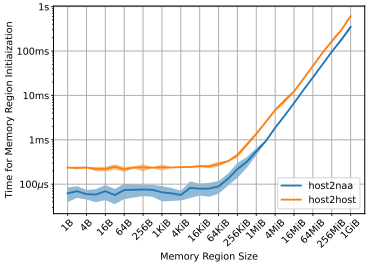


(c) Communication Bandwidth

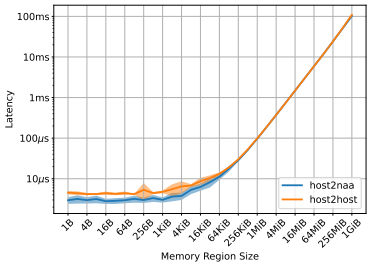
Performance Evaluation

Evaluation of RDMA communication with NAA (FPGA framework)

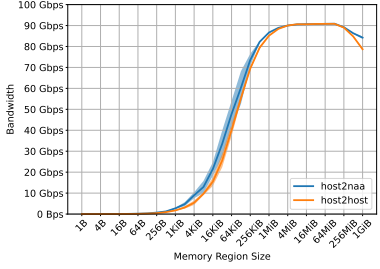
- Time for Memory Region Setup Protocol, latency, bandwidth; Switched 100 Gbps connection
- Baseline: host-to-host (software prototype of FPGA framework)



(a) Time for MRSP



(b) Communication Latency



(c) Communication Bandwidth

- Performance very close to theoretical maximum
- More details in SC'23 H2RC workshop paper: *Enabling Communication with FPGA-based Network-attached Accelerators for HPC Workloads* (to appear)

HPC Center Integration

Goal 2: Integration into processes and infrastructure of ...

1. Users:
   - Job scripts
   - Application-dependent Reconfiguration
   - Reporting of job's energy usage on NAA

2. Datacenter operations
   - Interaction with and support for resource management system and memory management
   - NAA system integration $\rightarrow$ power control, status monitoring
   - Energy monitoring

Need for Energy Monitoring $\rightarrow$ Goal 3

Summary

- Project's goal: enable flexible and scalable usage of network-attached FPGAs in HPC context
- So far: successfully demonstrated efficient communication
- Integration into workflows and infrastructure crucial for success

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

This project (grant no: 16ME0622K, 16ME0623) is sponsored by the Federal Ministry of Education and Research.

Project website: greenhpc.eu

Summary

- Project's goal: enable flexible and scalable usage of network-attached FPGAs in HPC context
- So far: successfully demonstrated efficient communication
- Integration into workflows and infrastructure crucial for success

Questions? Discussion!

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

This project (grant no: 16ME0622K, 16ME0623) is sponsored by the Federal Ministry of Education and Research.

Project website: greenhpc.eu