TosKonnect

A Modular Queue-based Communication Layer for Heterogeneous High Performance Computing

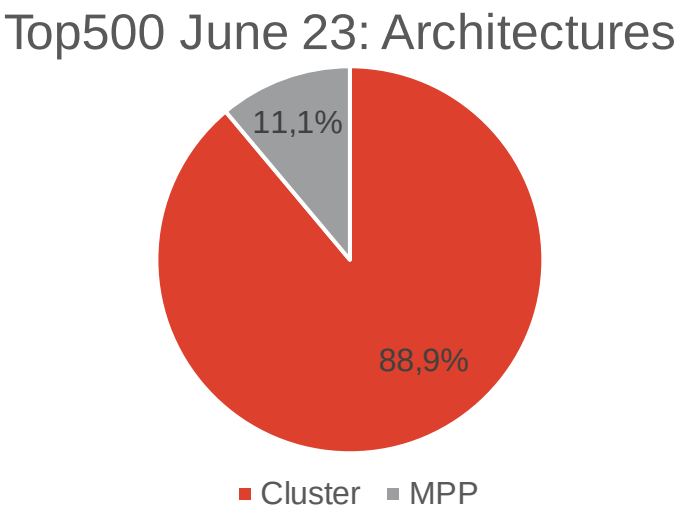
Laura Fuentes Grau

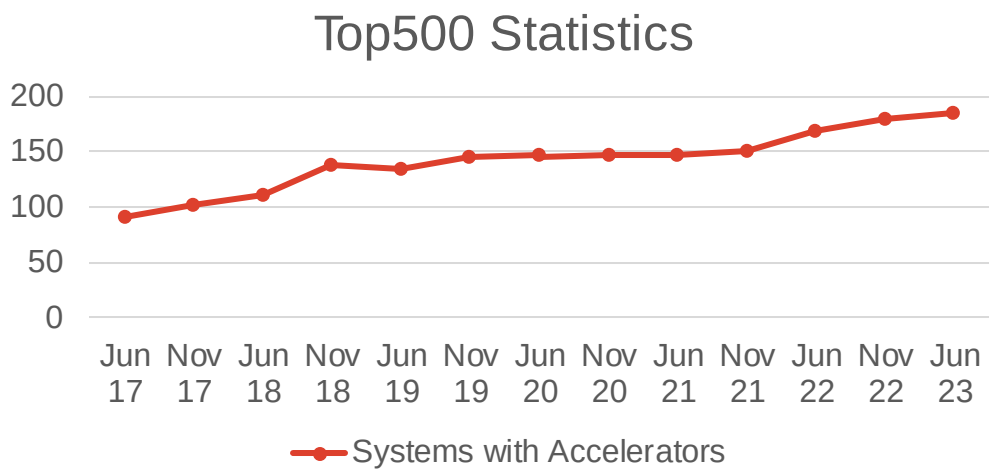
ACS | Automation of Complex Power Systems

E.ON Energy Research Center

RWTH AACHEN UNIVERSITY

The Problem





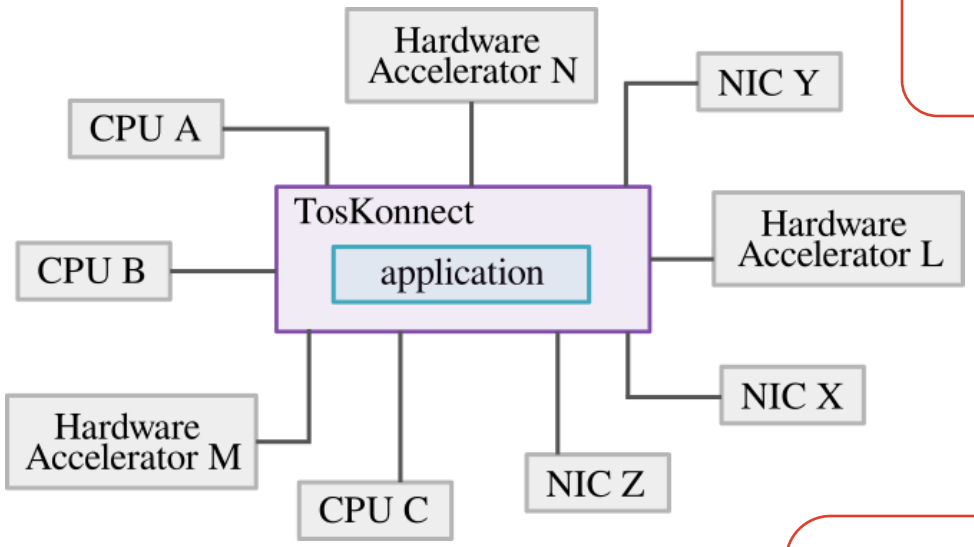
- Too high-level or complex
  - Possible use case: lower level of existing protocol stacks (e.g., MPI)
- Only focused on inter-machine networking
- Little diversity in communication methods
  - Focus on asynchronous queue-based communication

The Creation of TosKonnect

Vendor-neutral and Device-independent

Scalable and portable

High bandwidth data transfers



Queue-based communication

Flexible and extensible device support

Concepts

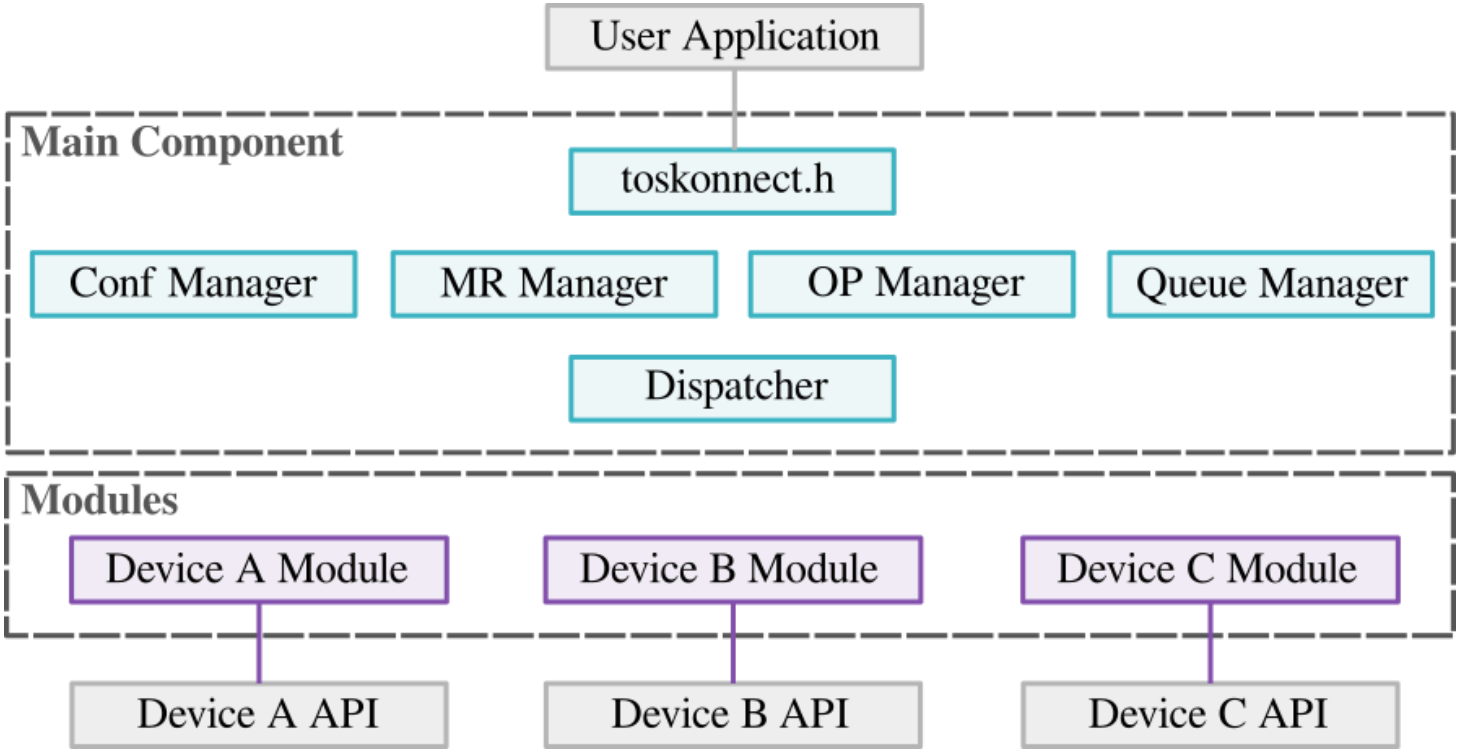
Modularity
to support flexibility and extensibility

Abstraction
to unify and configure data transfers

Asynchronous data transfers
to enable latency hiding

Queue-based communication
to streamline data transfer organization

Modularity: Software Architecture



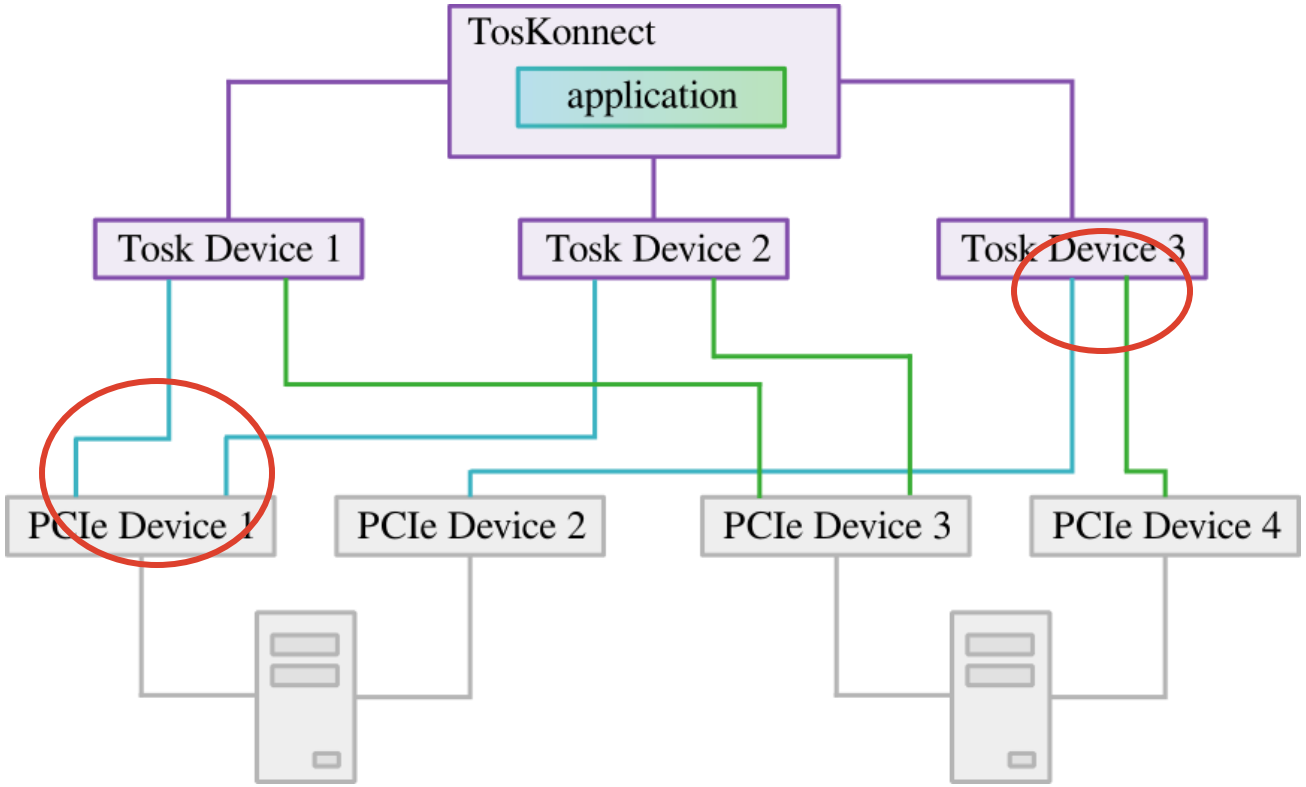
Modularity
to support flexibility and extensibility

Abstraction
to unify and configure data transfers

Asynchronous data transfers
to enable latency hiding

Queue-based communication
to streamline data transfer organization

Abstraction: TosKonnect devices



- A TosKonnect device is a communication endpoint

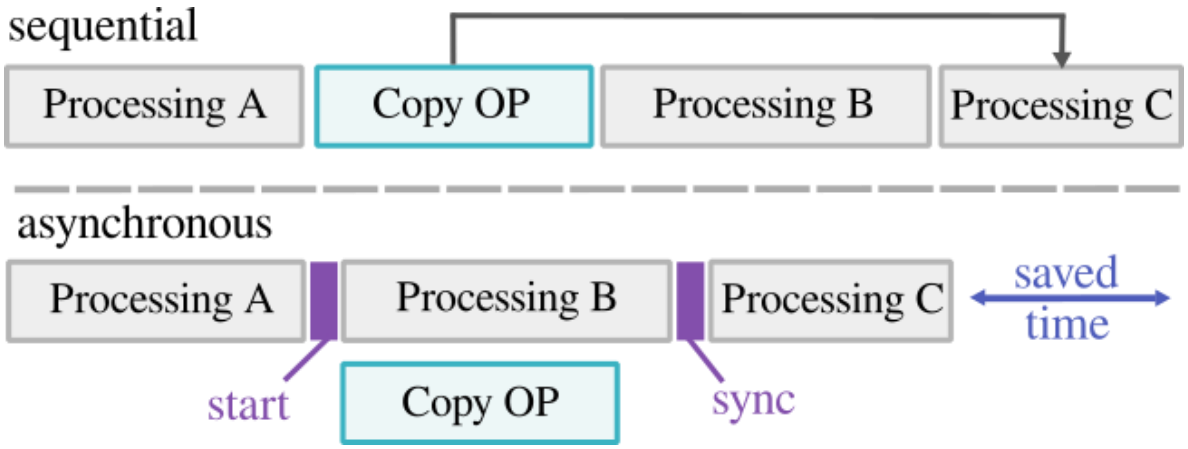
Modularity
to support flexibility and extensibility

Abstraction
to unify and configure data transfers

Asynchronous data transfers
to enable latency hiding

Queue-based communication
to streamline data transfer organization

Asynchronous data transfers



- Overlap processing and data transfers
  - ≡ hide data transfer latency
- Lower execution time compared to sequential version

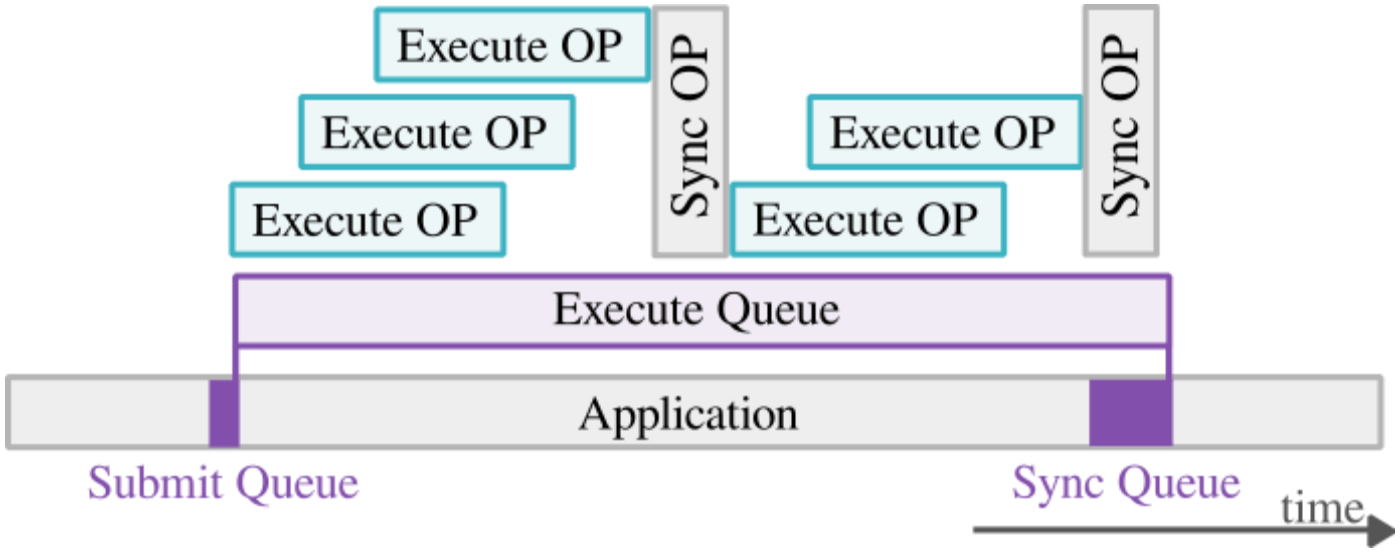
Modularity
to support flexibility and extensibility

Abstraction
to unify and configure data transfers

Asynchronous data transfers
to enable latency hiding

Queue-based communication
to streamline data transfer organization

Asynchronous Queue-Based Communication I



- TosKonnect executes each queue in a separate thread
- Each operation executes asynchronously
- Sync OPs create synchronization barriers
  - Execution stops until all previous operations are complete

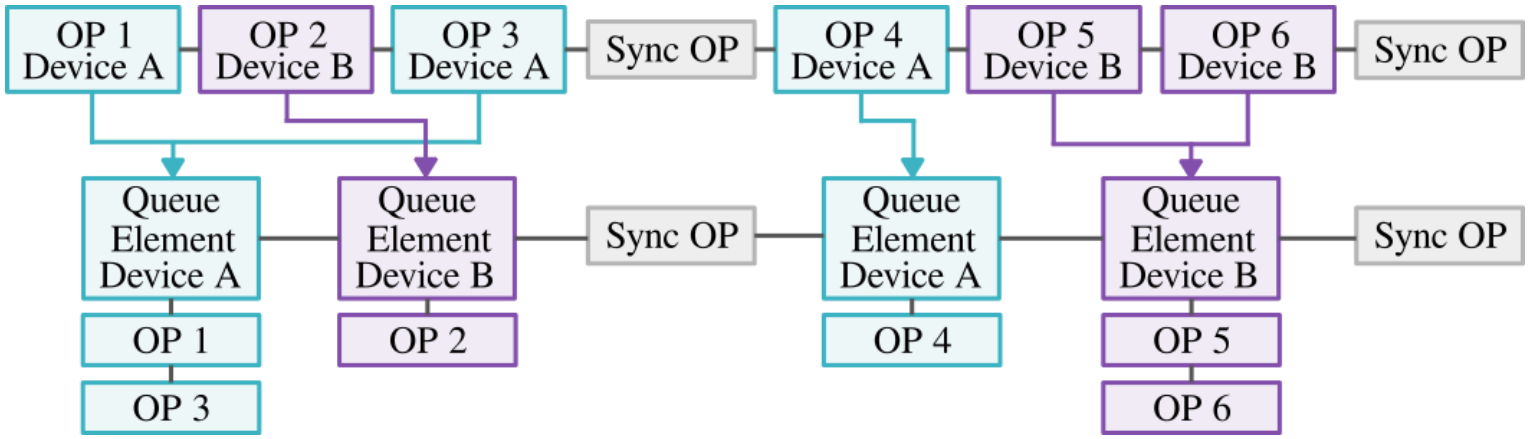
Modularity
to support flexibility and extensibility

Abstraction
to unify and configure data transfers

Asynchronous data transfers
to enable latency hiding

Queue-based communication
to streamline data transfer organization

Queue-Based Communication II



- Operations are executed by the modules
- Operations for the same device in between barriers are grouped into queue elements

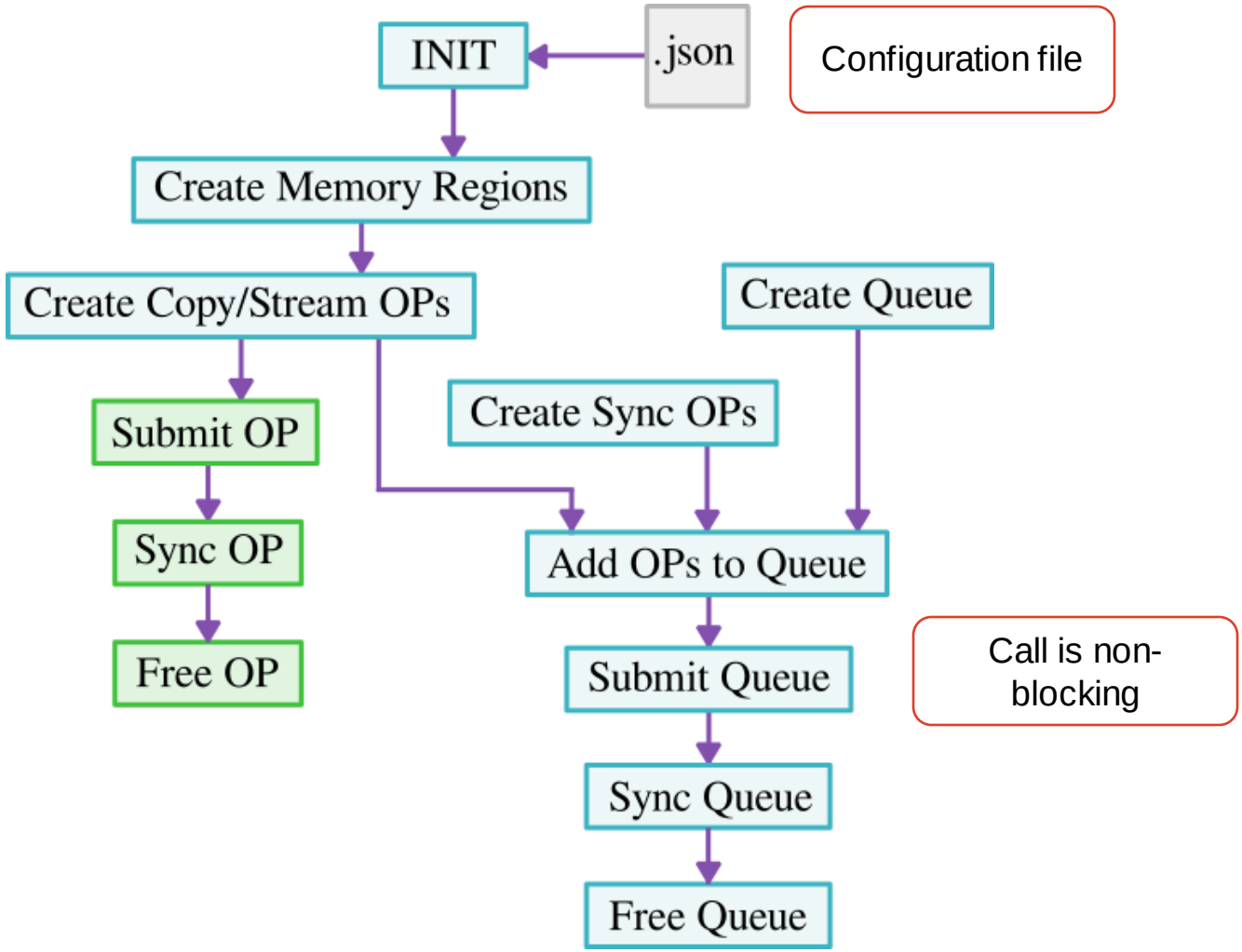
Modularity
to support flexibility and extensibility

Abstraction
to unify and configure data transfers

Asynchronous data transfers
to enable latency hiding

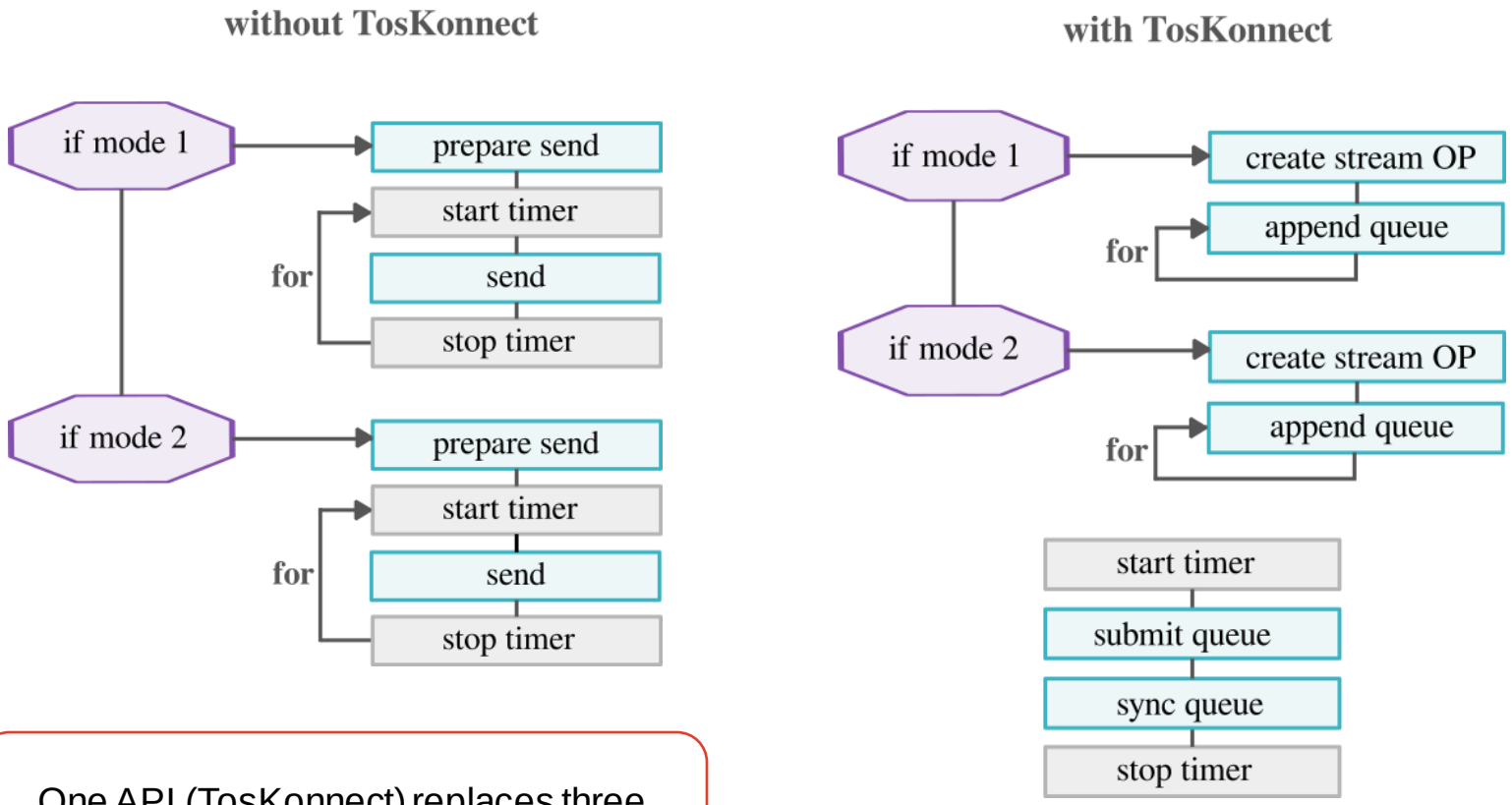
Queue-based communication
to streamline data transfer organization

API lifecycle



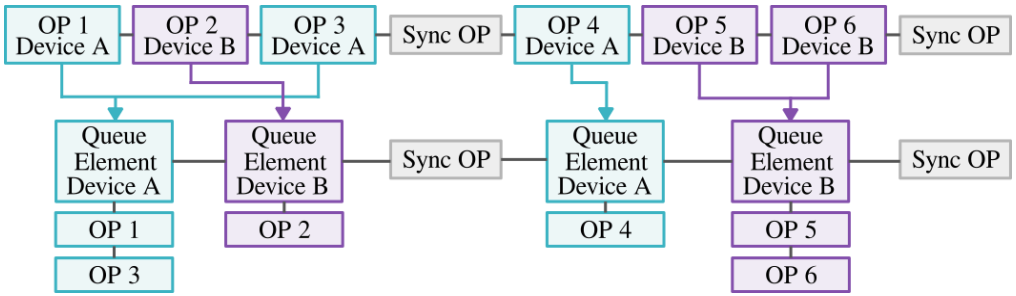
Example: InfiniGPUDirect

- Benchmark application for GPUDirect RDMA with InfiniBand

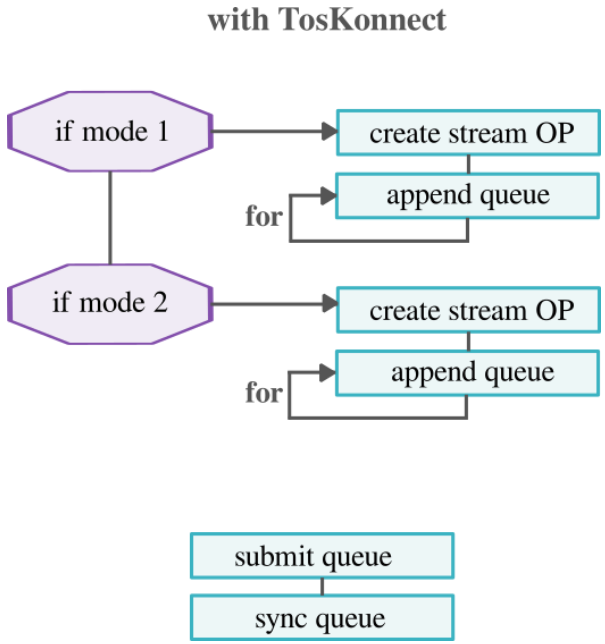


One API (TosKonnect) replaces three (CUDA, InfiniBand, TCP)

InfiniGPUDirect: TosKonnect timer

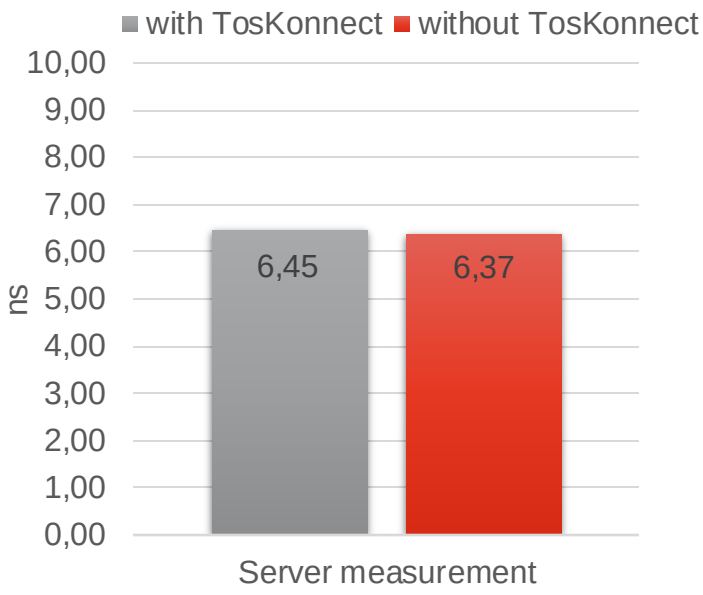


- Measure time in the queue thread
  - In between sync objects
- No need for external timers
- Measurement closer to pure copy calls

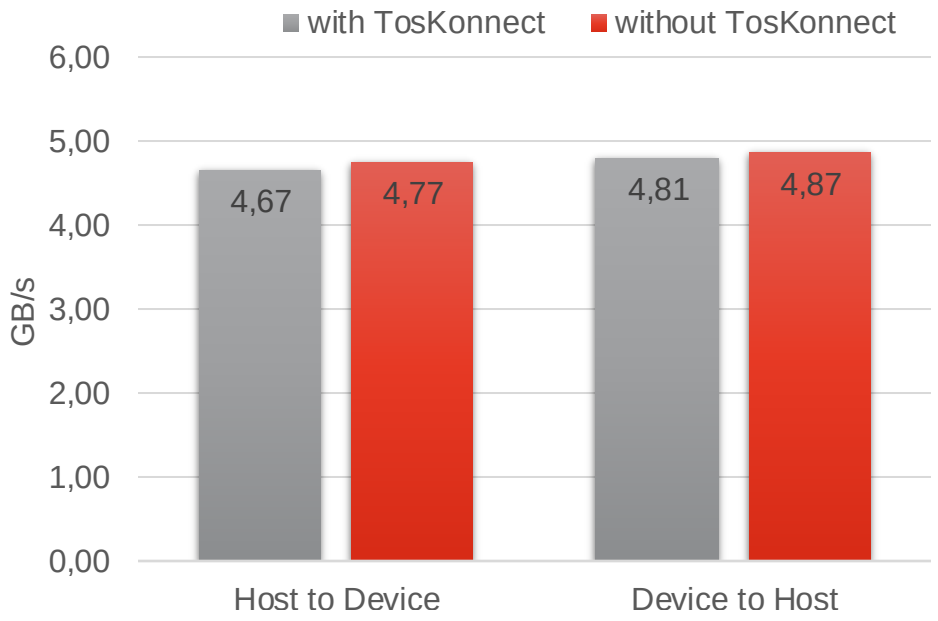


InfiniGPUDirect: Evaluation

Execution time

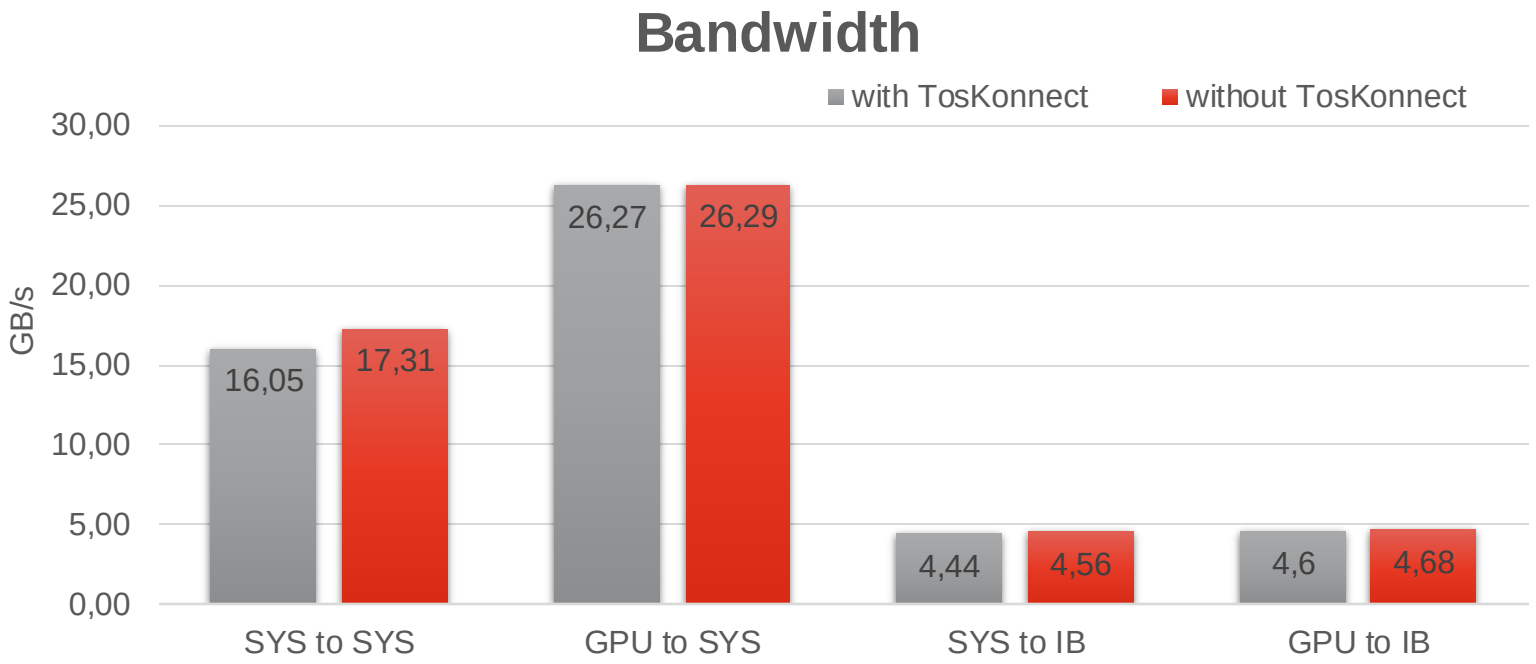


Bandwidth



- Approx. 1% execution time overhead
- Approx. 2% less bandwidth
- TosKonnect decouples InfiniBand and GPU device APIs
  - Application code shrinks to less than 50% of the original size

Further Benchmarks



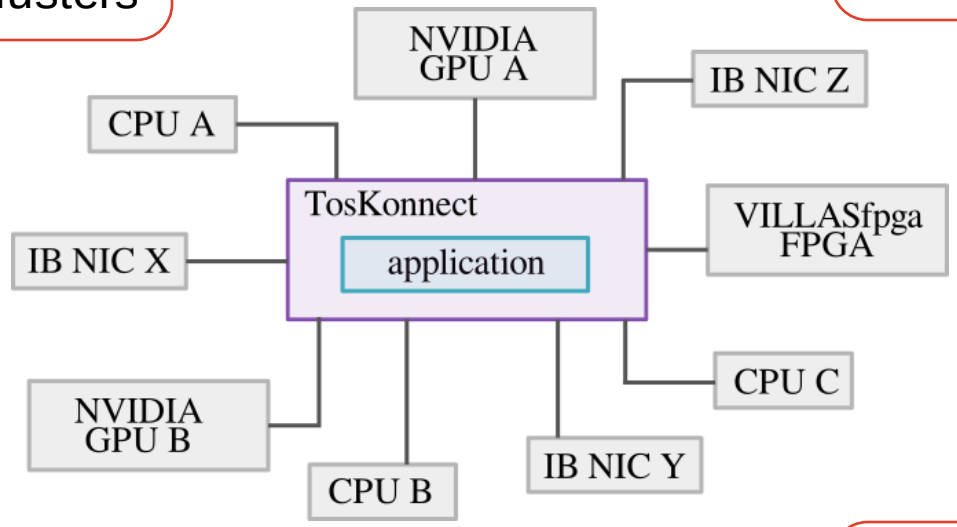
- Low overhead
- Overhead is dominated by thread creation
  - Necessary for asynchronous nature

Conclusion

Viable communication layer for heterogeneous clusters

Allows streamlined code restructure

Up to 50% less boilerplate code



Enables latency hiding

Scalable and extensible

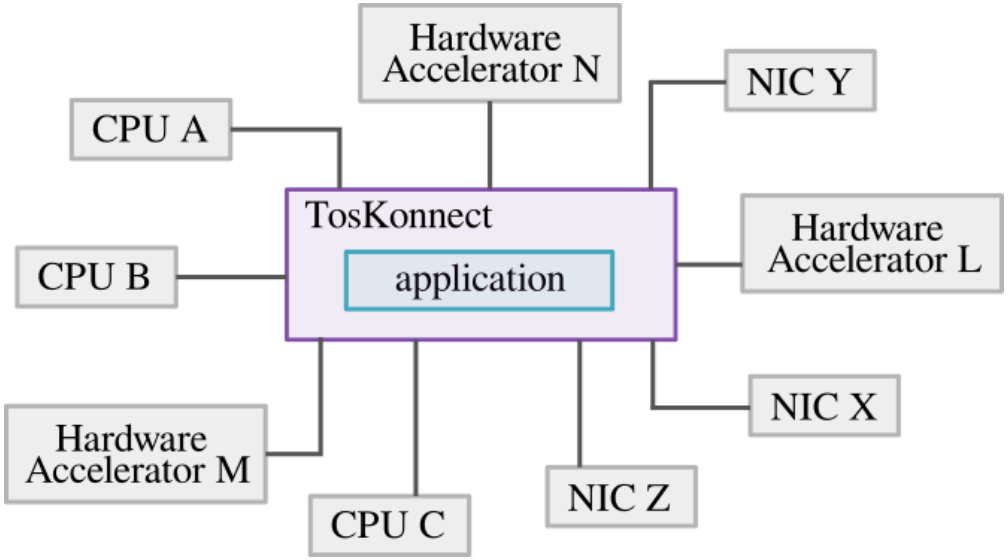
Integrated support to compare device performance

Supports interoperability

Introduces only a small amount of overhead

Future work

- Active encouragement of TosKonnect integration and usage
- Include advanced functions
- Offer more device support
  - e.g. VILLASfpga



Contact

E.ON Energy Research Center
Mathieustraße 10
52074 Aachen
Germany

Laura Fuentes Grau
laura.fuentes-grau@eonerc.rwth-aachen.de
http://www.eonerc.rwth-aachen.de

ACS | Automation of Complex Power Systems

E.ON Energy Research Center

RWTH AACHEN UNIVERSITY