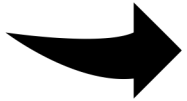
SmartNIC-driven Accelerator-centric Architecture for Network Servers

Maroun Tork Lina Maudlej Mark Silberstein

ACSL Lab
Technion, Israel

Agenda



- Motivation and Background
- Design
- Evaluation
- Conclusions

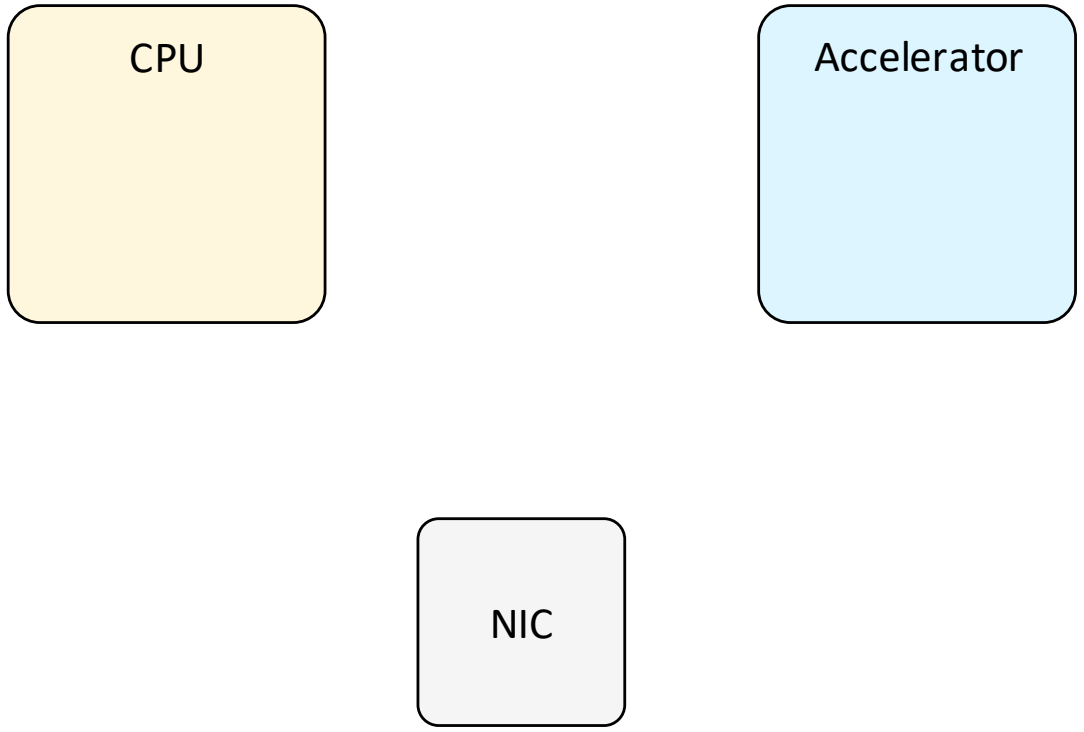
Data center trends:
AI workloads require hardware acceleration

Data center trends:
AI workloads require hardware acceleration

Data center trends:
AI workloads require hardware acceleration

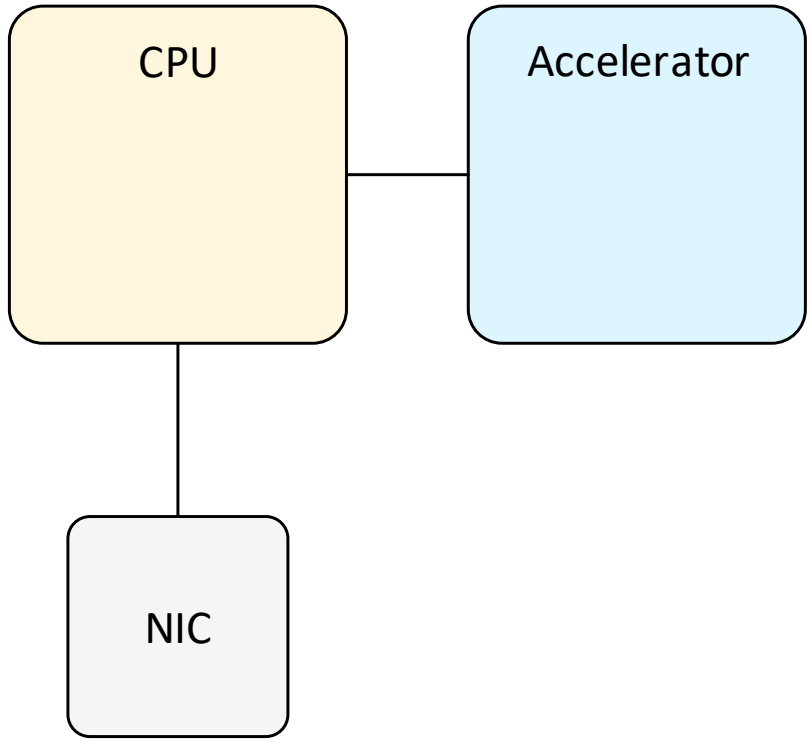
Accelerated
Network
Servers

Accelerated Server Architectures



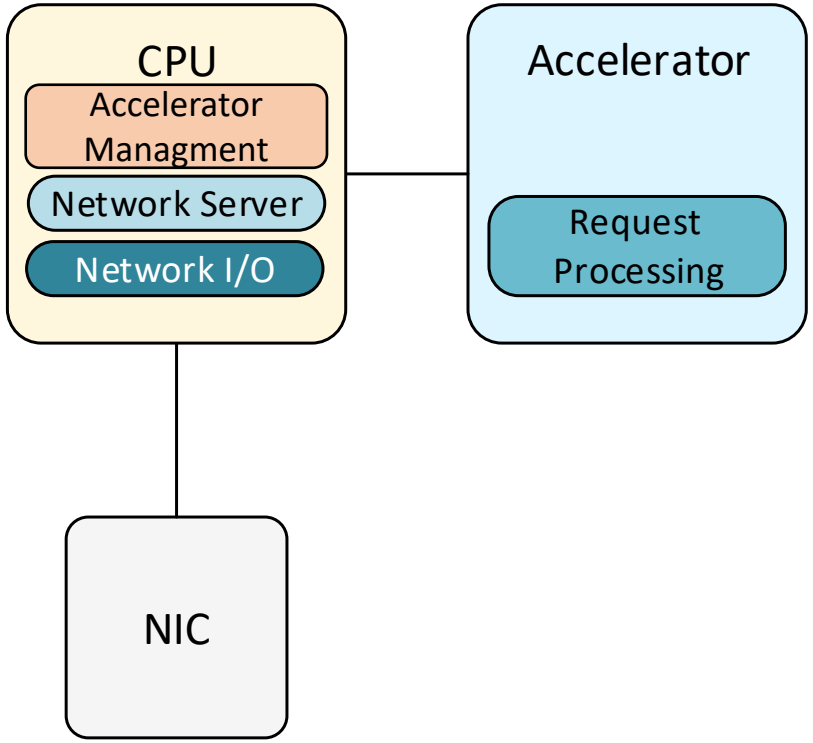
Accelerated Server Architectures

Traditional host-centric



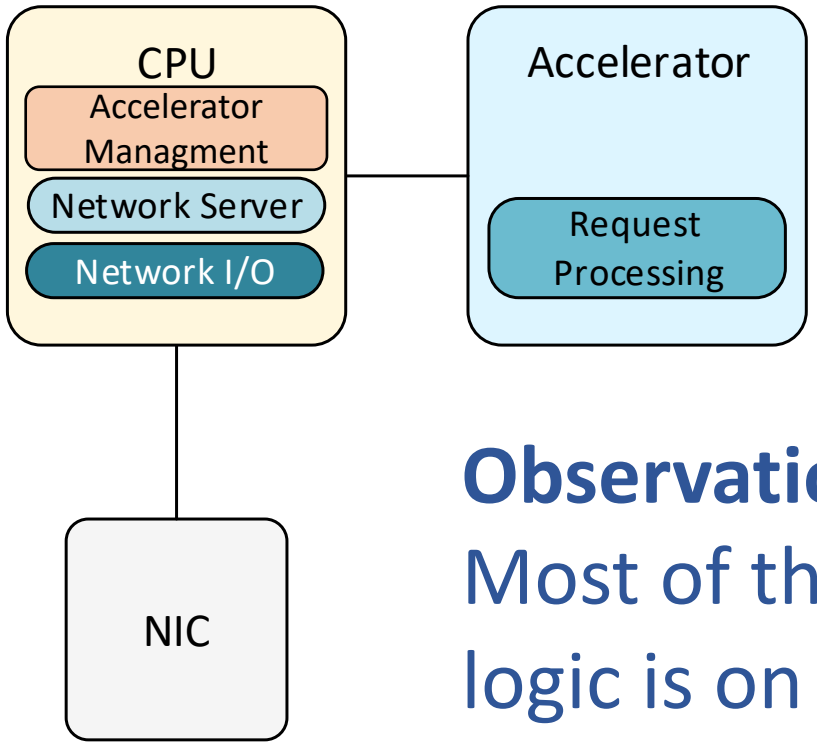
Accelerated Server Architectures

Traditional host-centric



Accelerated Server Architectures

Traditional host-centric



Observation

Most of the application-specific logic is on accelerator!

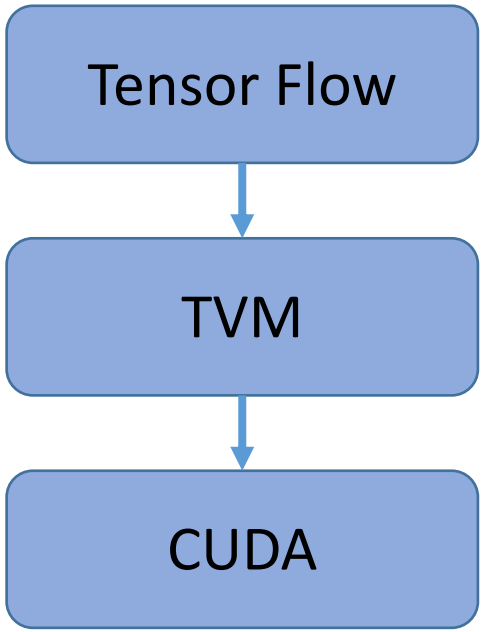
Full application offloading to accelerators

- Eliding CPU involvement to save interaction
  - Accelerator invocation, synchronization, data transfers.

Full application offloading to accelerators

- Eliding CPU involvement to save interaction
  - Accelerator invocation, synchronization, data transfers.

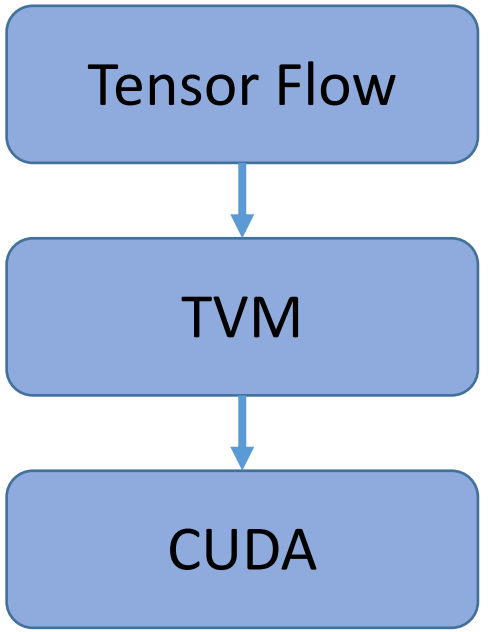
- Support by TVM
  - TVM compiles out CPU code
  - Kernel fusion to save data movements



Full application offloading to accelerators

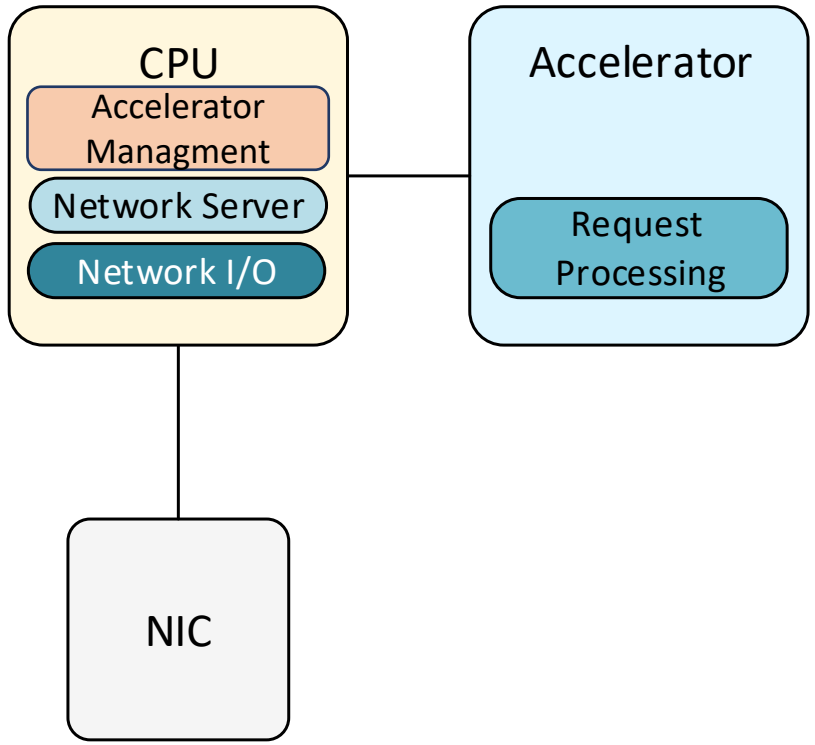
- Eliding CPU involvement to save interaction
  - Accelerator invocation, synchronization, data transfers.

- Support by TVM
  - TVM compiles out CPU code
  - Kernel fusion to save data movements



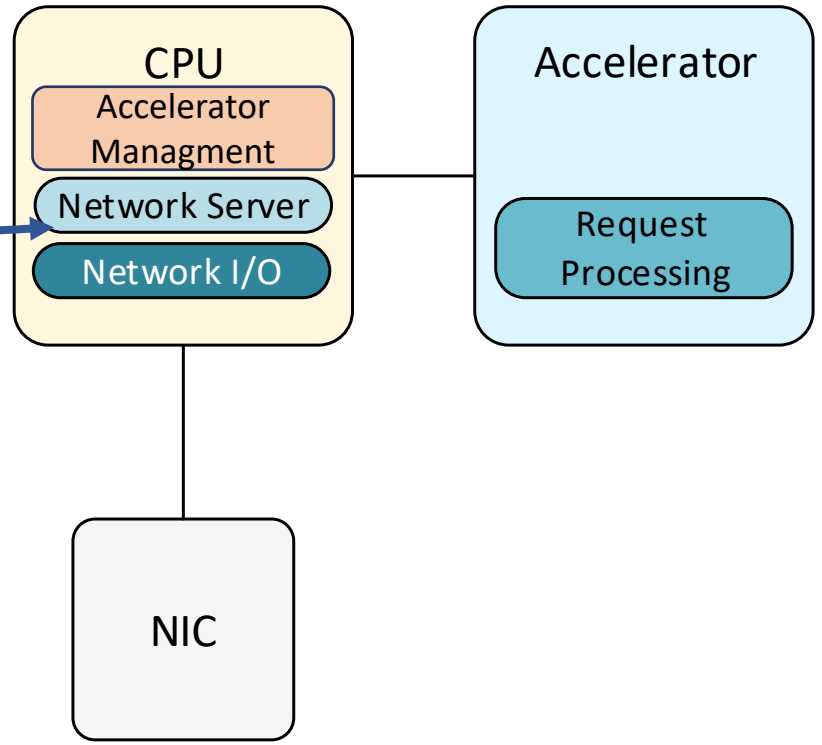
So why do we need a CPU?

CPU role in accelerated servers: I/O and accelerator management



CPU role in accelerated servers: I/O and accelerator management

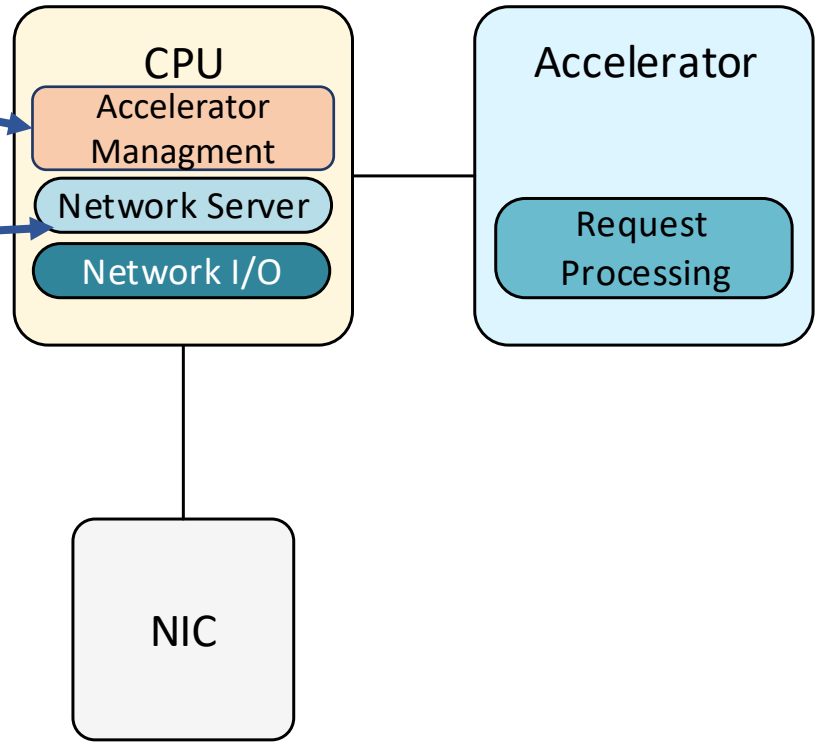
CPU runs network processing



CPU role in accelerated servers: I/O and accelerator management

CPU feeds the accelerator

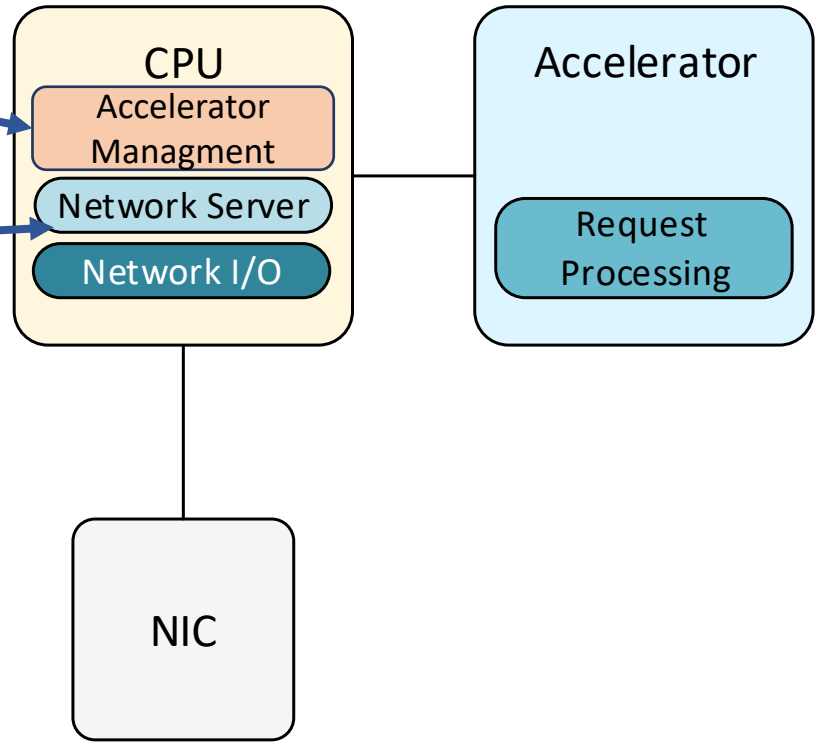
CPU runs network processing



CPU role in accelerated servers: I/O and accelerator management

CPU feeds the accelerator

CPU runs network processing



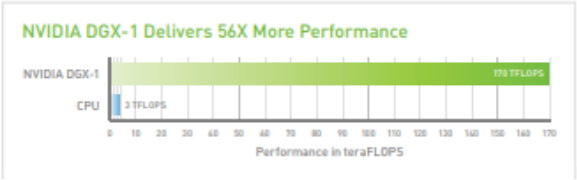
How many CPU are needed for these tasks?

Real world example: NVIDIA DGX-1

The World's First Deep Learning Supercomputer in a Box

Data scientists and artificial intelligence (AI) researchers require accuracy, simplicity, and speed for deep learning success. Faster training and iteration ultimately means faster innovation and time-to-market.

The NVIDIA® DGX-1™ is the world's first purpose-built system optimized for deep learning, with fully integrated hardware and software that can be deployed quickly and easily. Its revolutionary performance significantly accelerates training time, making it the world's first deep learning supercomputer in a box.



SYSTEM SPECIFICATIONS

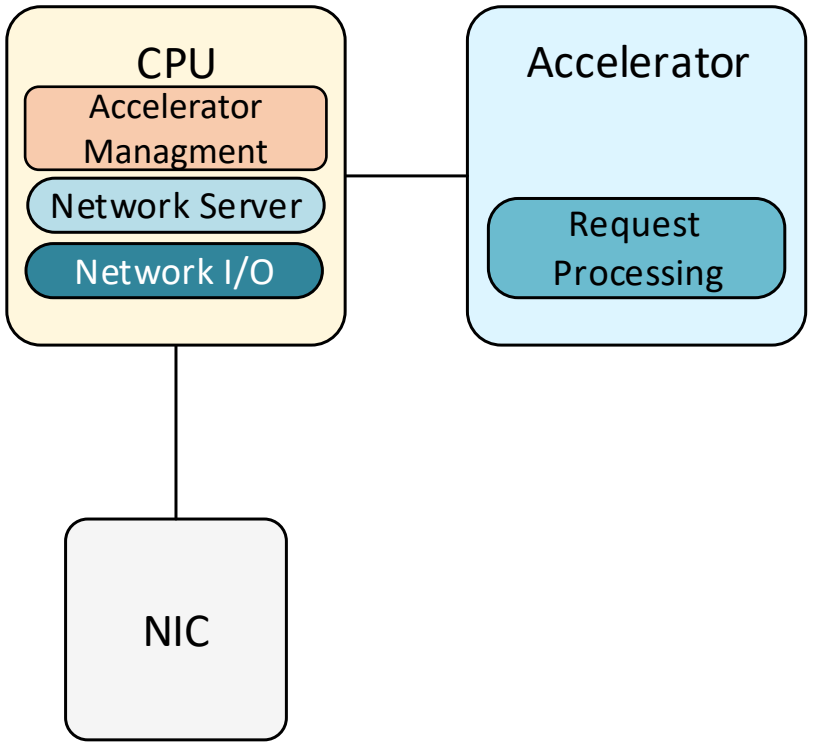
GPUs	8x Tesla GP100
TFLOPS (GPU FP16 / CPU FP32)	170/3
GPU Memory	16 GB per GPU
CPU	Dual 20-core Intel® Xeon® E5-2698 v4 2.2 GHz
NVIDIA CUDA® Cores	28672
System Memory	512 GB 2133 MHz DDR4 LRDIMM
Storage	4x 1.92 TB SSD RAID 0
Network	Dual 10 GbE, 4 IB EDR
Software	Ubuntu Server Linux OS DGX-1 Recommended GPU Driver
System Weight	134 lbs
System Dimensions	866 D x 444 W x 131 H (mm)
Packing Dimensions	1180 D x 730 W x 284 H (mm)
Maximum Power Requirements	3200W
Operating Temperature Range	10 - 35 °C

SYSTEM SPECIFICATIONS

GPUs	8x Tesla GP100
TFLOPS (GPU FP16 / CPU FP32)	170/3
GPU Memory	16 GB per GPU
CPU	Dual 20-core Intel® Xeon® E5-2698 v4 2.2 GHz
NVIDIA CUDA® Cores	28672
System Memory	512 GB 2133 MHz DDR4 LRDIMM
Storage	4x 1.92 TB SSD RAID 0
Network	Dual 10 GbE, 4 IB EDR
Software	Ubuntu Server Linux OS DGX-1 Recommended GPU Driver
System Weight	134 lbs
System Dimensions	866 D x 444 W x 131 H (mm)
Packing Dimensions	1180 D x 730 W x 284 H (mm)
Maximum Power Requirements	3200W
Operating Temperature Range	10 - 35 °C

Accelerated Server Architectures

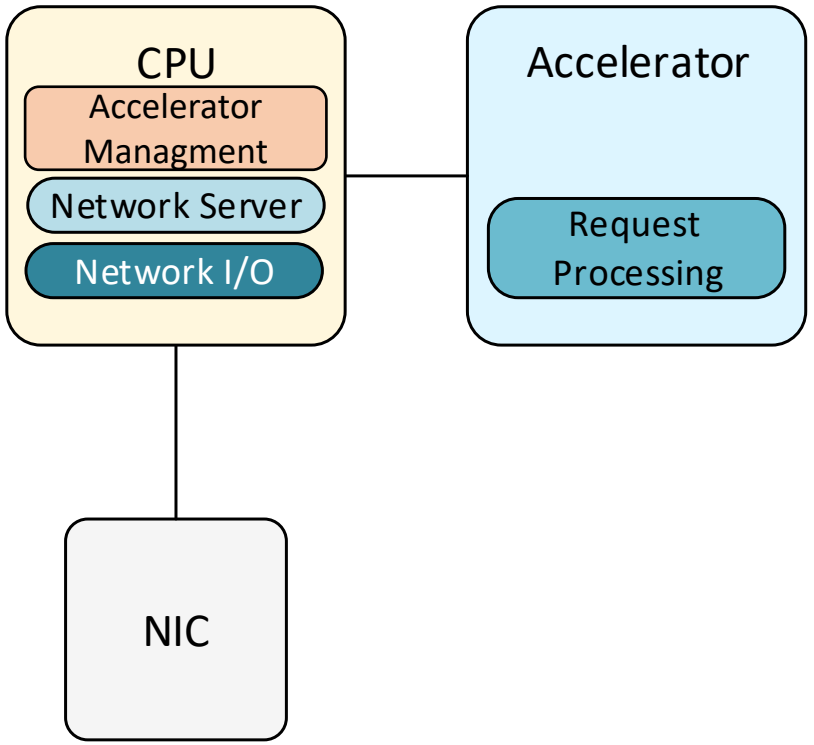
Traditional host-centric



Using Xeon core as an accelerator manager is inefficient:

Accelerated Server Architectures

Traditional host-centric

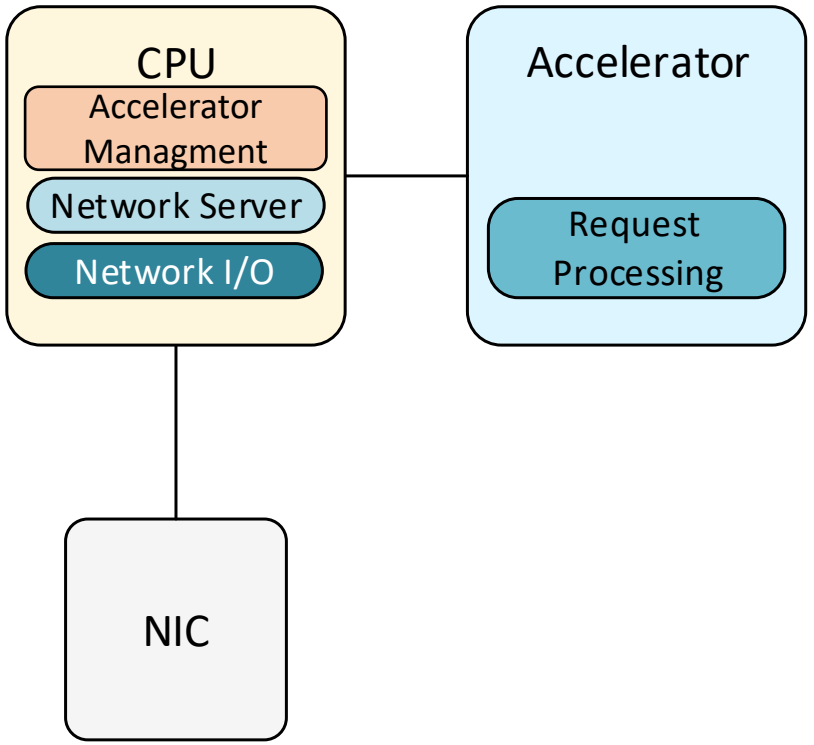


Using Xeon core as an accelerator manager is inefficient:

✗ Performance Bottleneck

Accelerated Server Architectures

Traditional host-centric

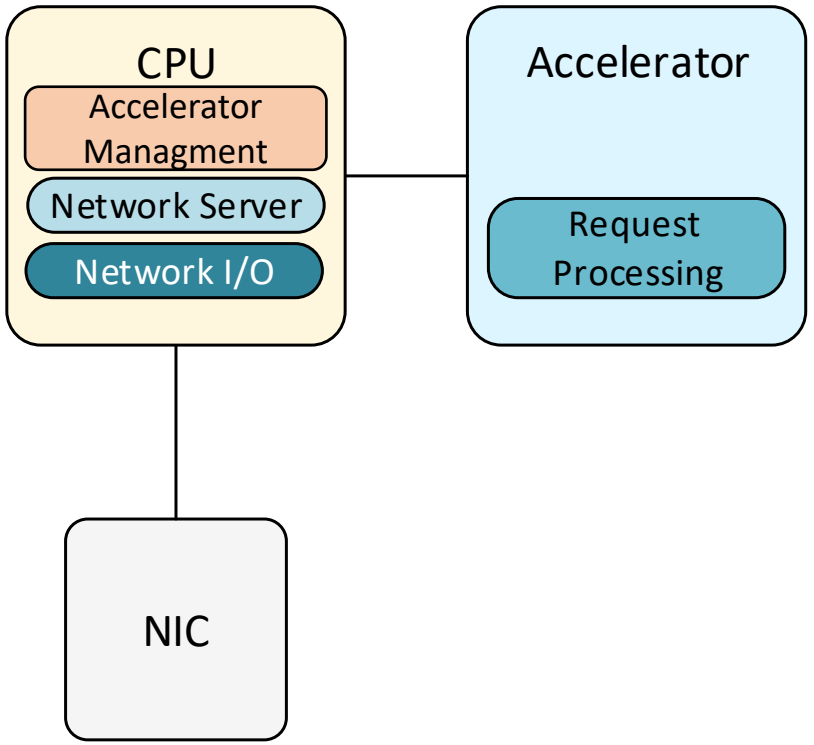


Using Xeon core as an accelerator manager is inefficient:

- ✗ Performance Bottleneck
- ✗ High cost
- ✗ Power consumption

Accelerated Server Architectures

Traditional host-centric

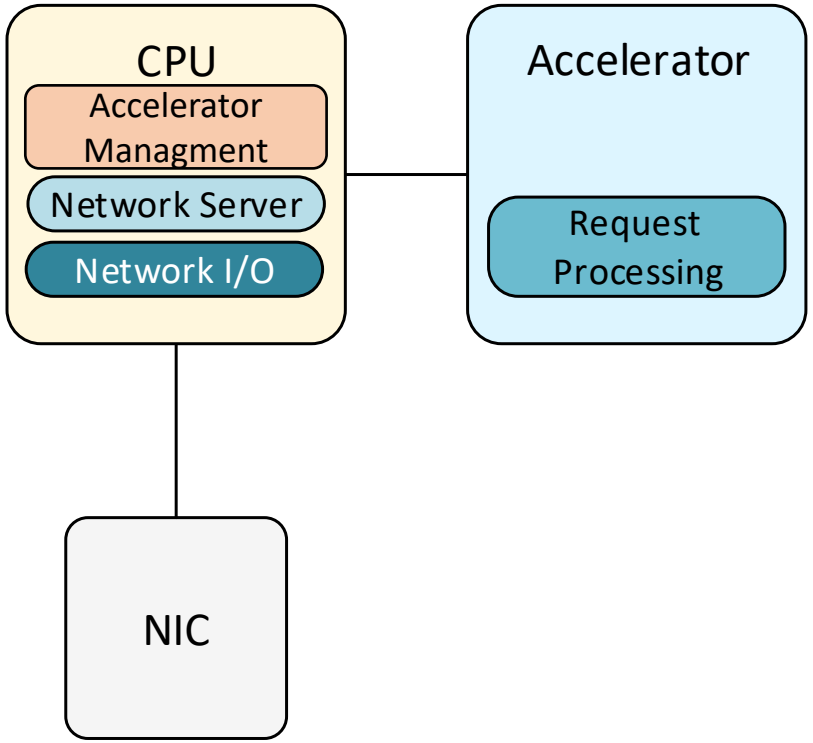


Using Xeon core as an accelerator manager is inefficient:

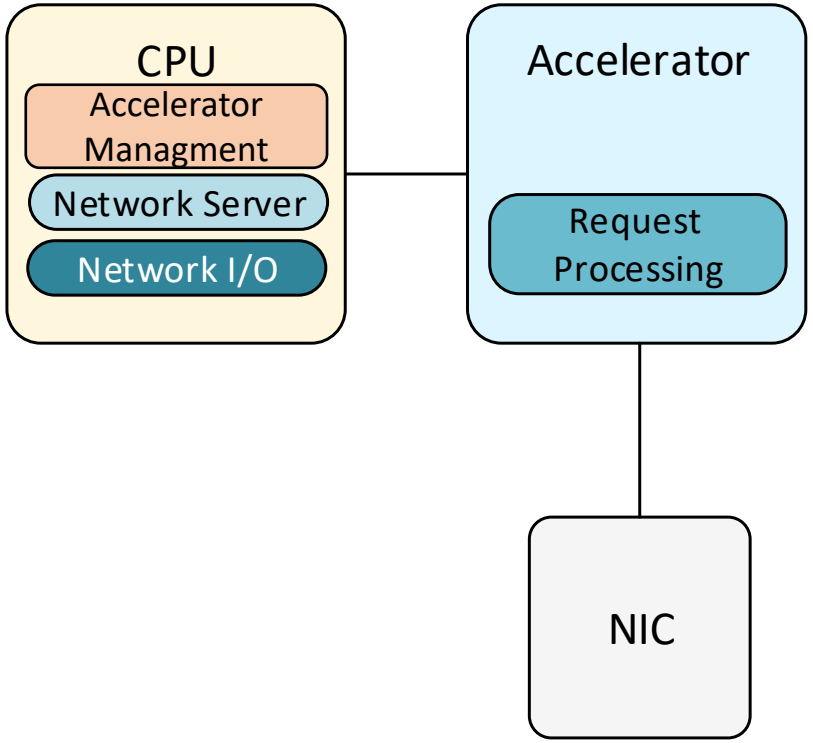
- ✗ Performance Bottleneck
- ✗ High cost
- ✗ Power consumption
- ✗ Tail latency due to noisy neighbors

Accelerated Server Architectures

Traditional host-centric

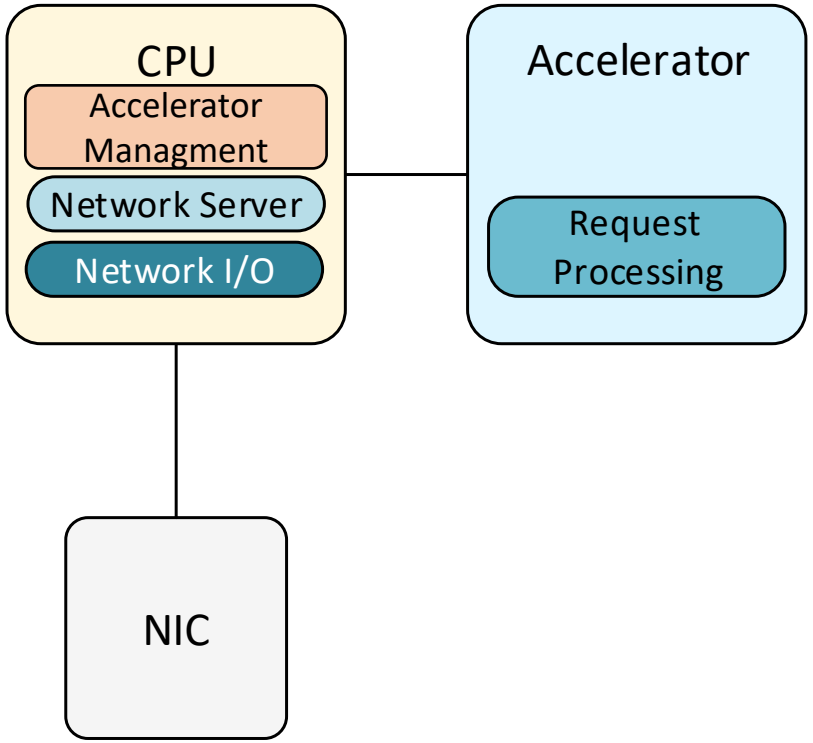


Accelerator-centric

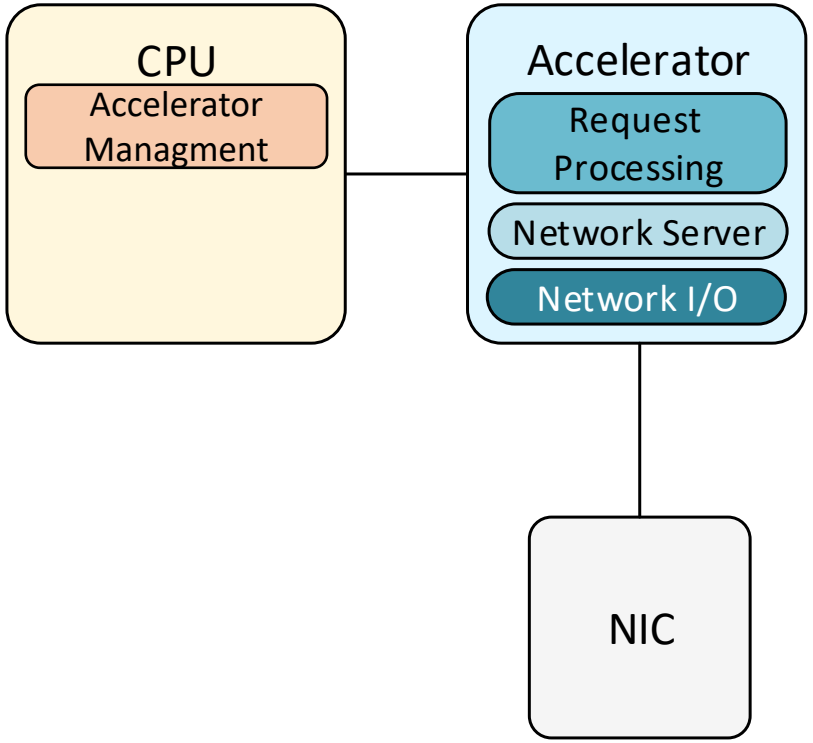


Accelerated Server Architectures

Traditional host-centric

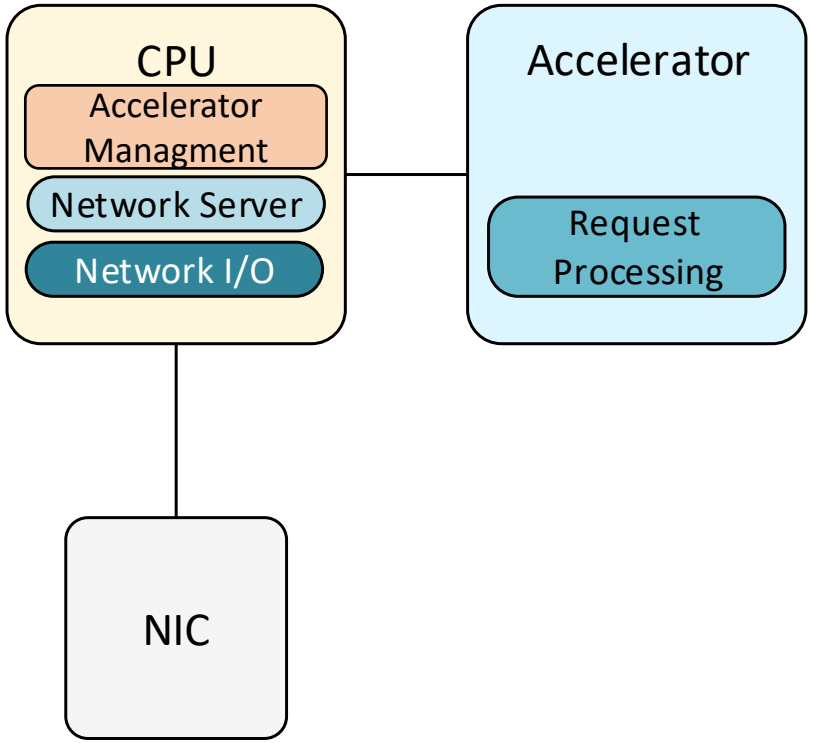


Accelerator-centric

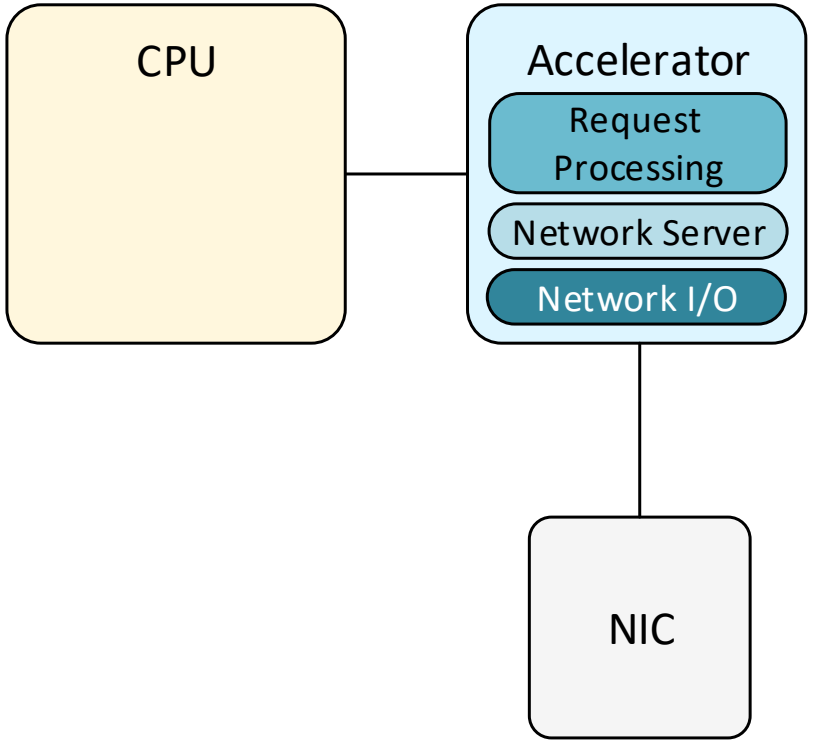


Accelerated Server Architectures

Traditional host-centric

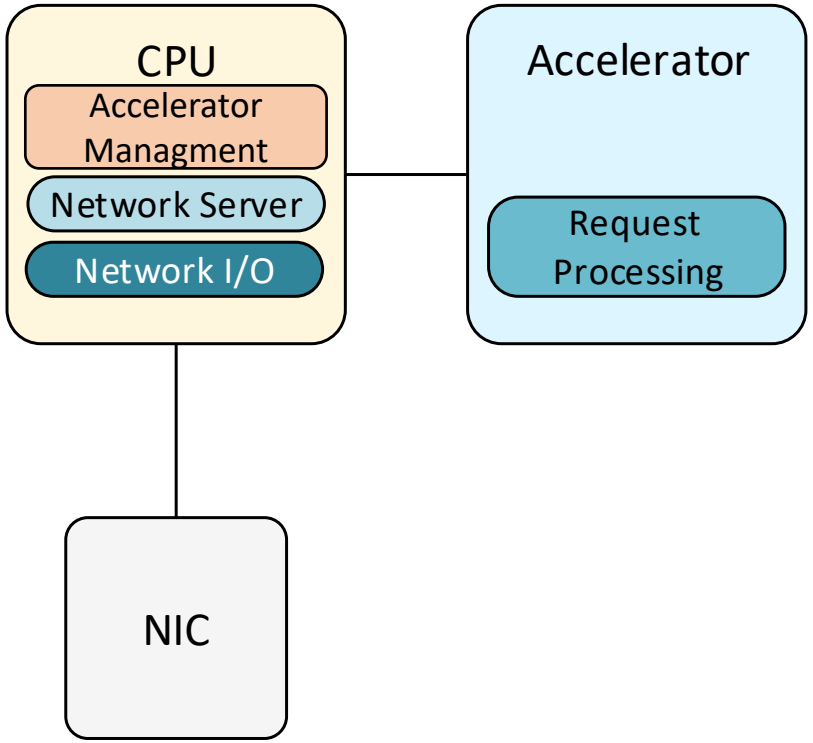


Accelerator-centric

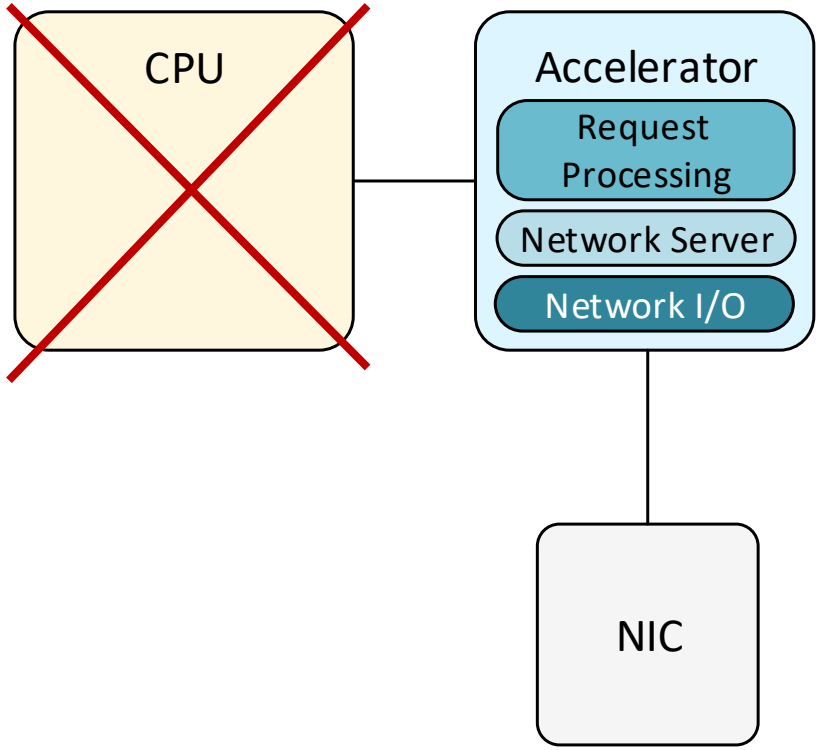


Accelerated Server Architectures

Traditional host-centric



Accelerator-centric

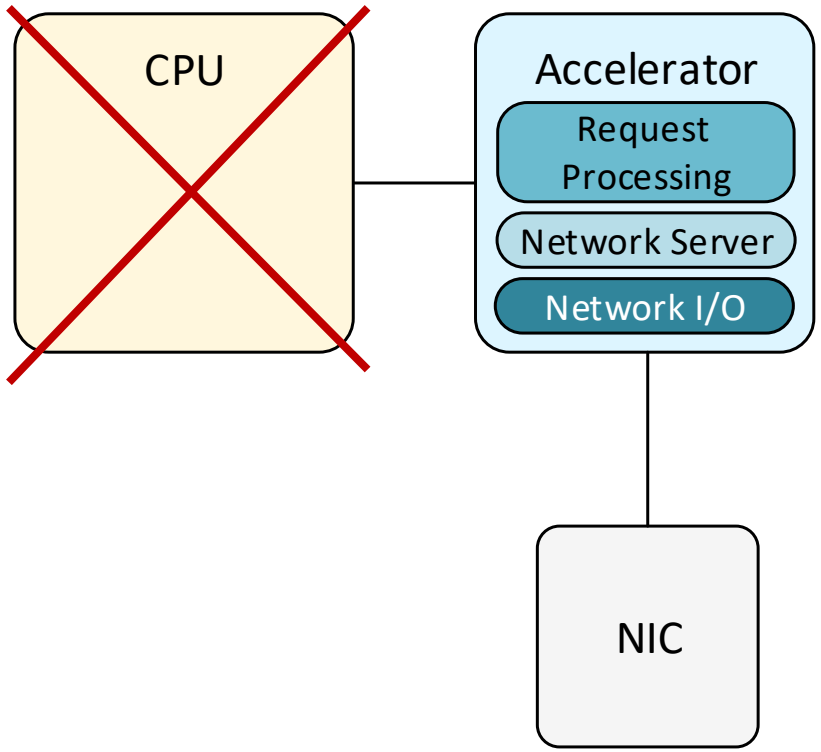


Accelerated Server Architectures

Previous works:

- IB-Verbs on GPUs – 2014
- GPUnet – 2014
- GPUrdma – 2016
- GPU-Centric – 2017

Accelerator-centric

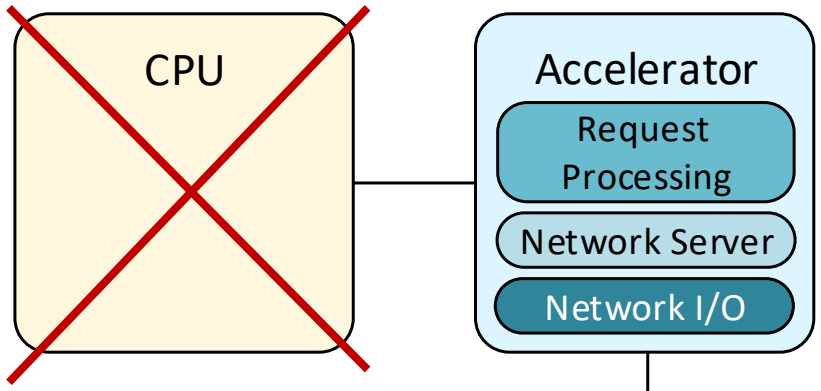


Accelerated Server Architectures

Previous works:

- IB-Verbs on GPUs – 2014
- GPUnet – 2014
- GPUrdma – 2016
- GPU-Centric – 2017

Accelerator-centric



Limitations

1. ~~GPU only~~

In this work

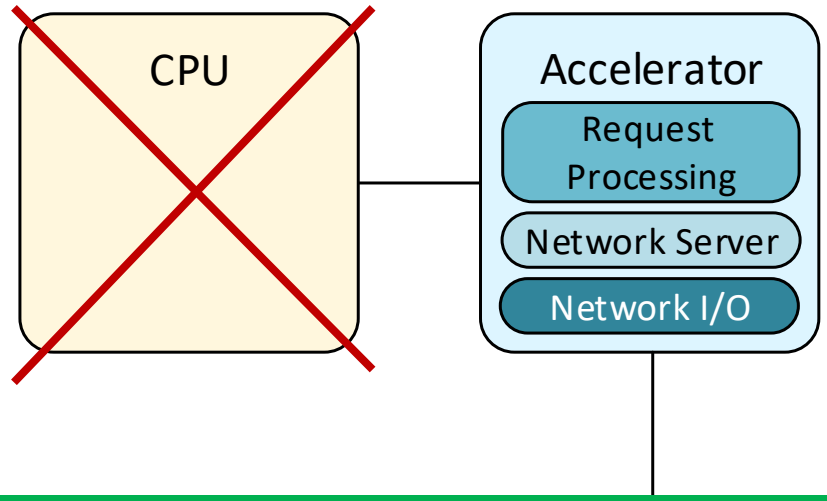
1. Portable

Accelerated Server Architectures

Previous works:

- IB-Verbs on GPUs – 2014
- GPUnet – 2014
- GPUrdma – 2016
- GPU-Centric – 2017

Accelerator-centric



Limitations

1. ~~GPU only~~
2. ~~InfiniBand only~~

In this work

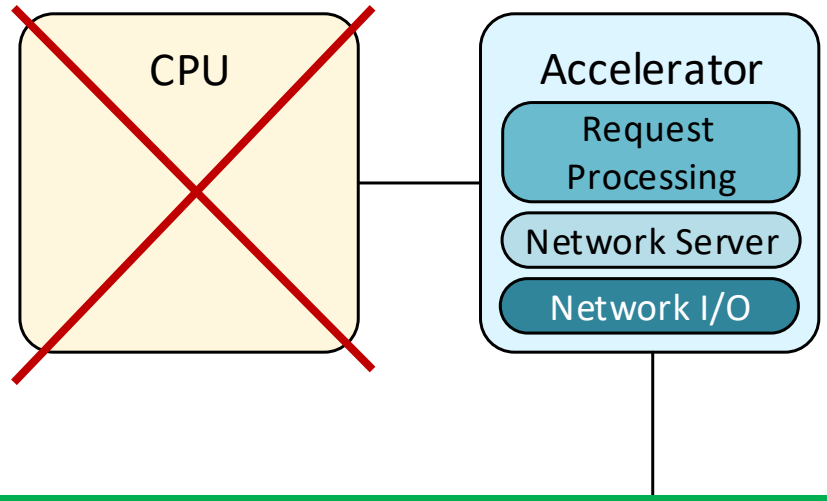
1. Portable
2. Support TCP/UDP

Accelerated Server Architectures

Previous works:

- IB-Verbs on GPUs – 2014
- GPUnet – 2014
- GPUrdma – 2016
- GPU-Centric – 2017

Accelerator-centric



Limitations

1. ~~GPU only~~
2. ~~InfiniBand only~~
3. ~~Accelerator Overhead~~

In this work

1. Portable
2. Support TCP/UDP
3. Offloaded server logic

Lynx - Vision

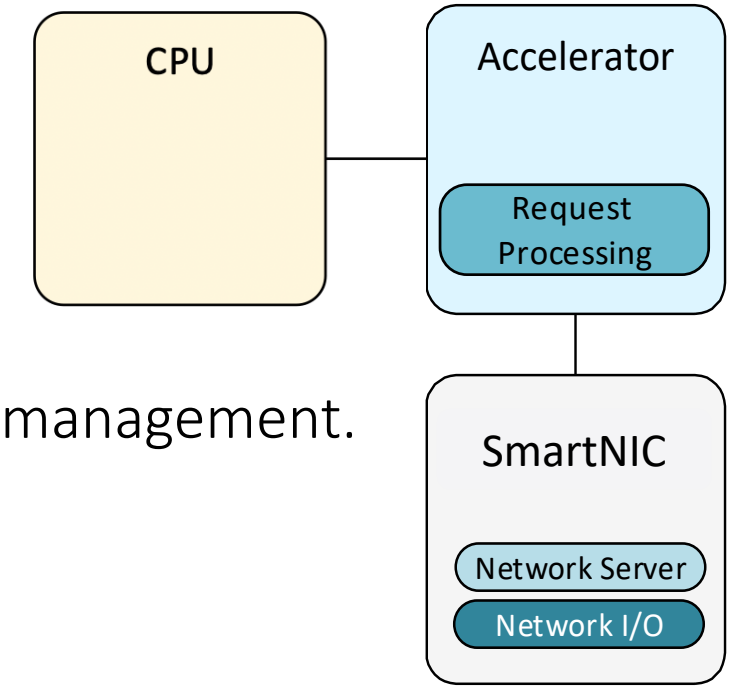
Goal

Demonstrate and build a general accelerated-centric server.

Lynx - Vision

Goal

Demonstrate and build a general accelerated-centric server.



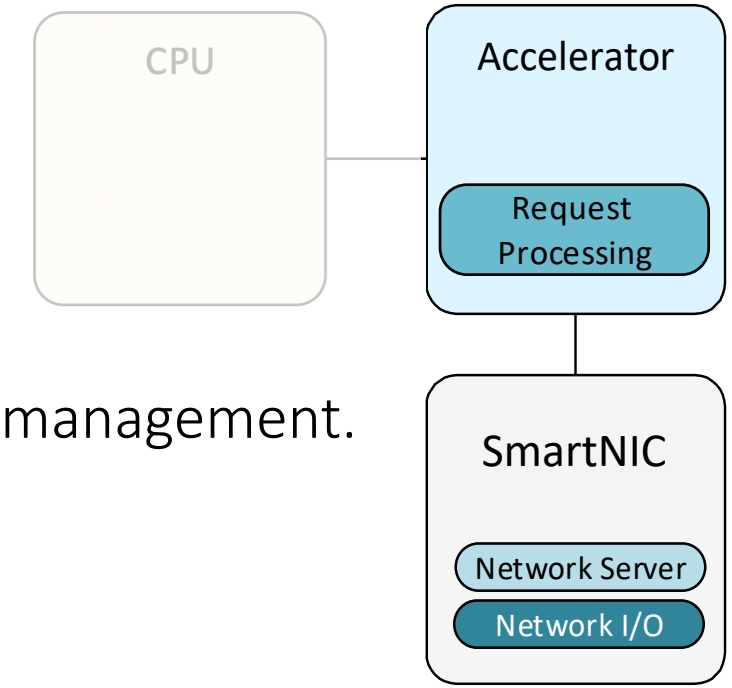
How?

Use SmartNIC for network processing and accelerator management.

Lynx - Vision

Goal

Demonstrate and build a general accelerated-centric server.

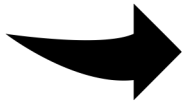


How?

Use SmartNIC for network processing and accelerator management.

- **Full CPU offload**
- **No application code on SmartNIC**

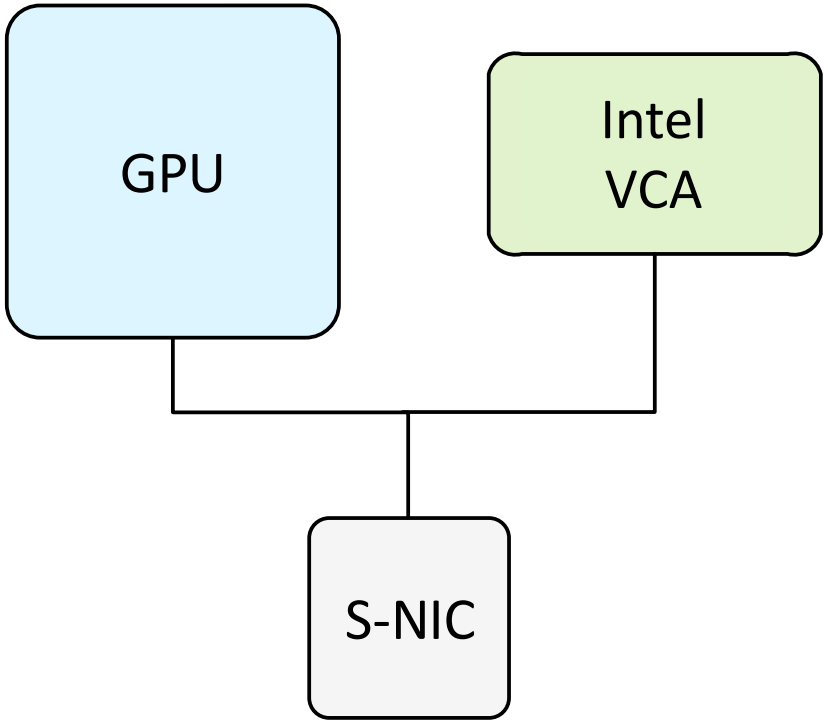
Agenda



- Motivation and Background
- Design
- Evaluation
- Conclusions

Challenges

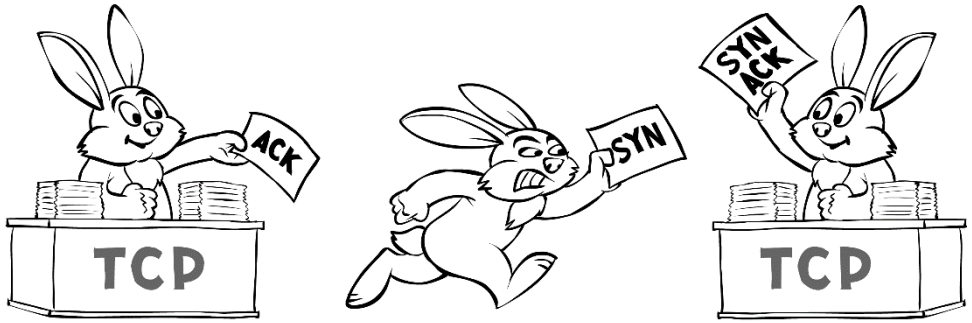
- Generic accelerator support.
  - Where is the driver?



Challenges

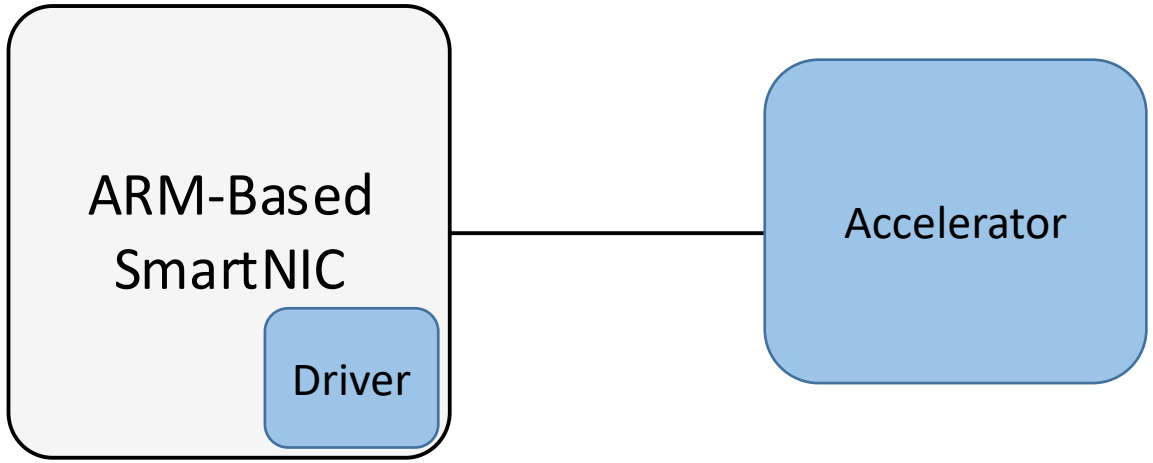
- Generic accelerator support

- Support for TCP/UDP transport protocols
  - Were is the network stack running?
    - CPU? No!



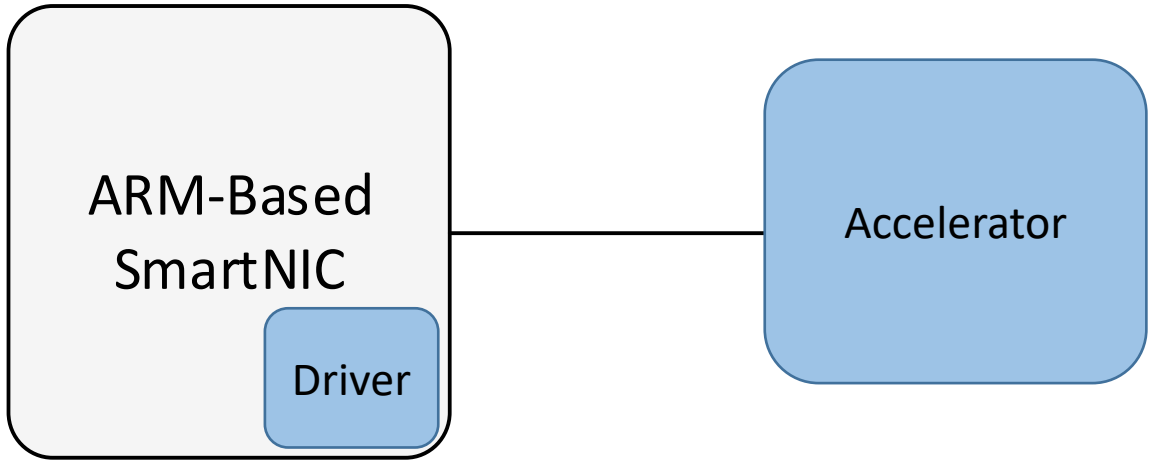
Design Principles

1. Portable Solution
   - Use Accelerator's DMA engine?



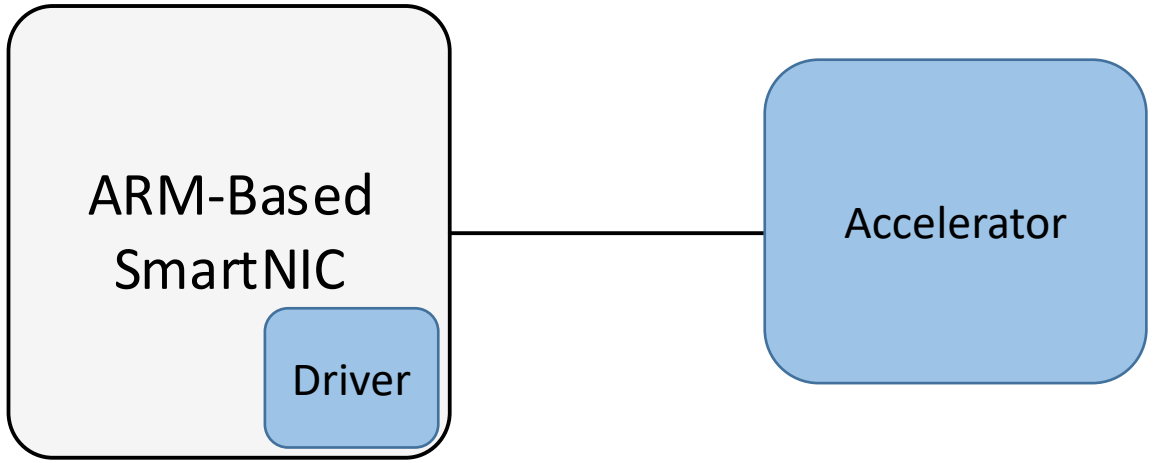
Design Principles

1. Portable Solution
   - Use Accelerator's DMA engine?
     We will need $K \cdot N$ drivers
     K – number of SmartNIC platforms
     N – number of accelerators

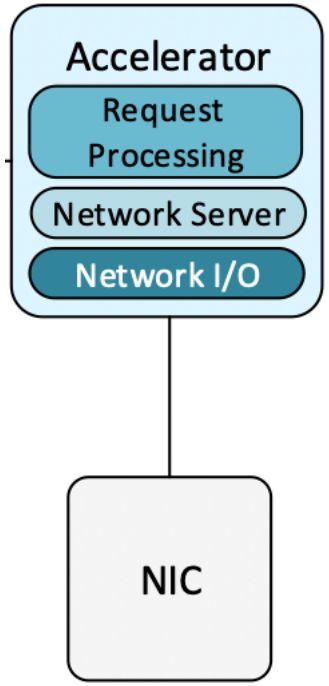


Design Principles

1. Portable Solution
   - Use Accelerator's DMA engine?
     We will need $K \cdot N$ drivers
     K – number of SmartNIC platforms
     N – number of accelerators

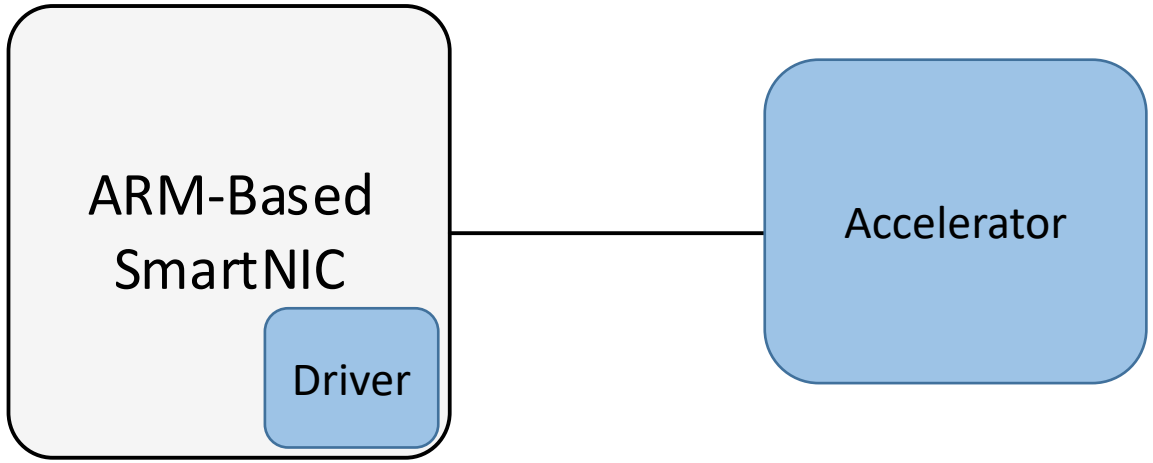


2. Network Processing
   - On the accelerator?

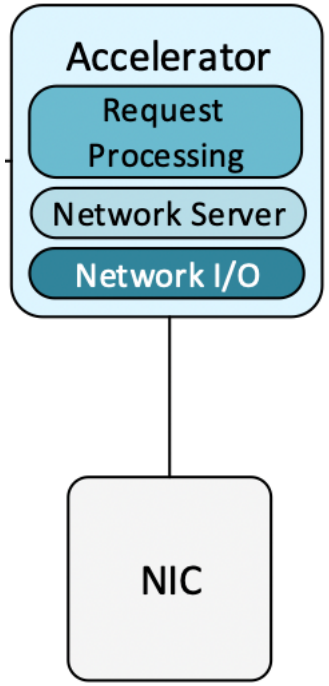


Design Principles

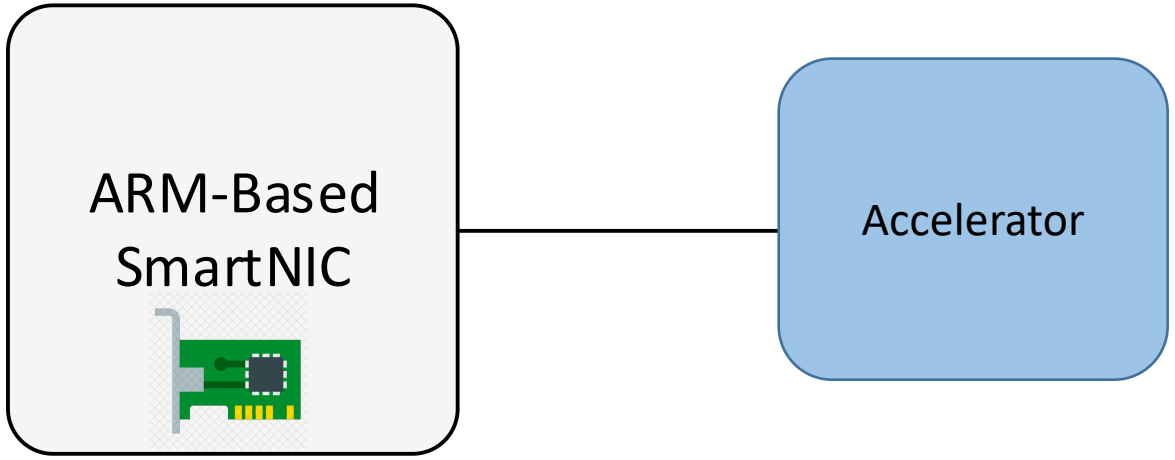
1. Portable Solution
   - Use Accelerator's DMA engine?
     We will need $K \cdot N$ drivers
     K – number of SmartNIC platforms
     N – number of accelerators



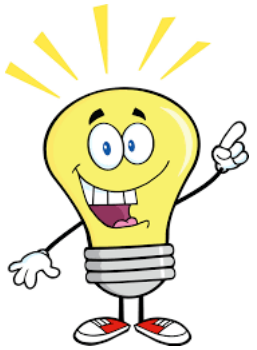
2. Network Processing
   - On the accelerator?
     Extra overhead when managing the NIC from the accelerator



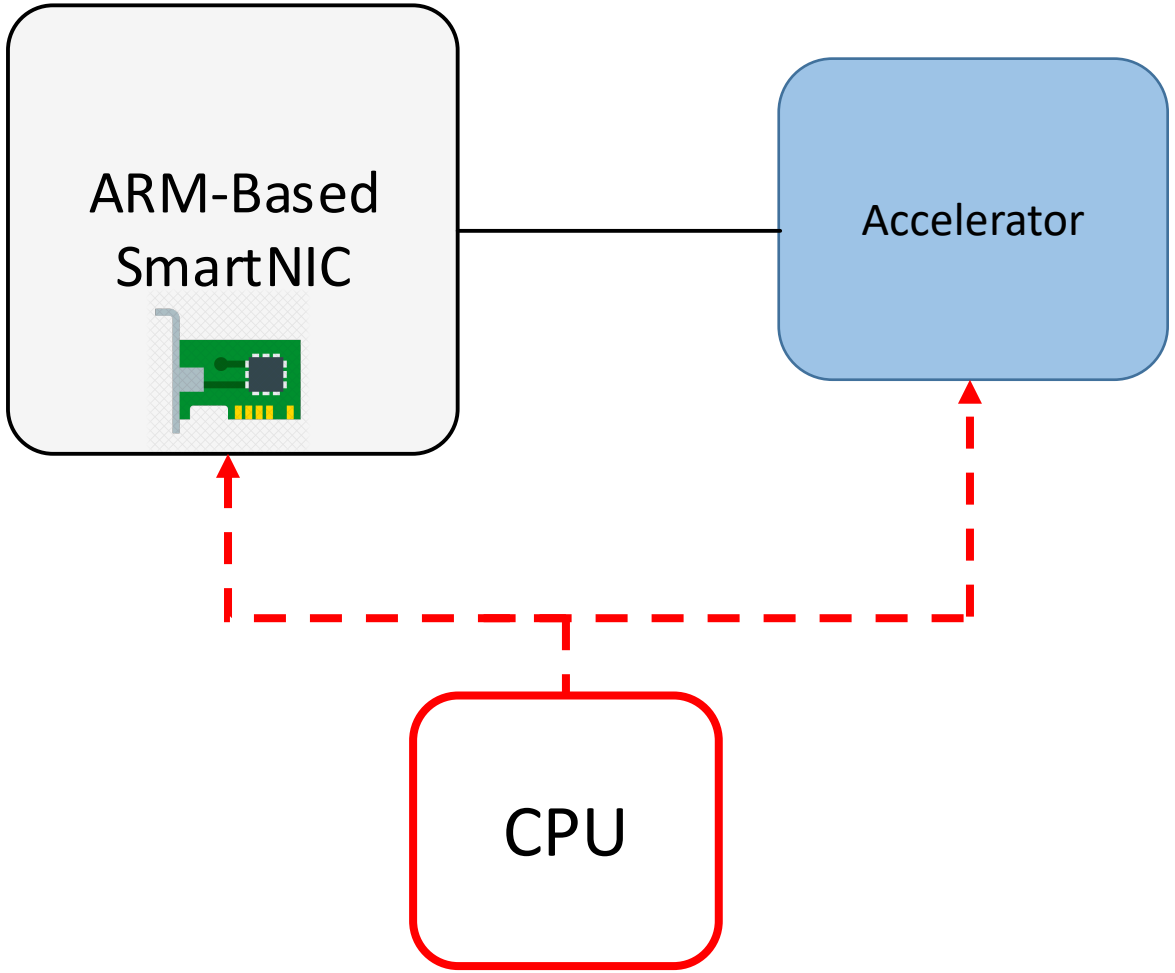
How can we do it?



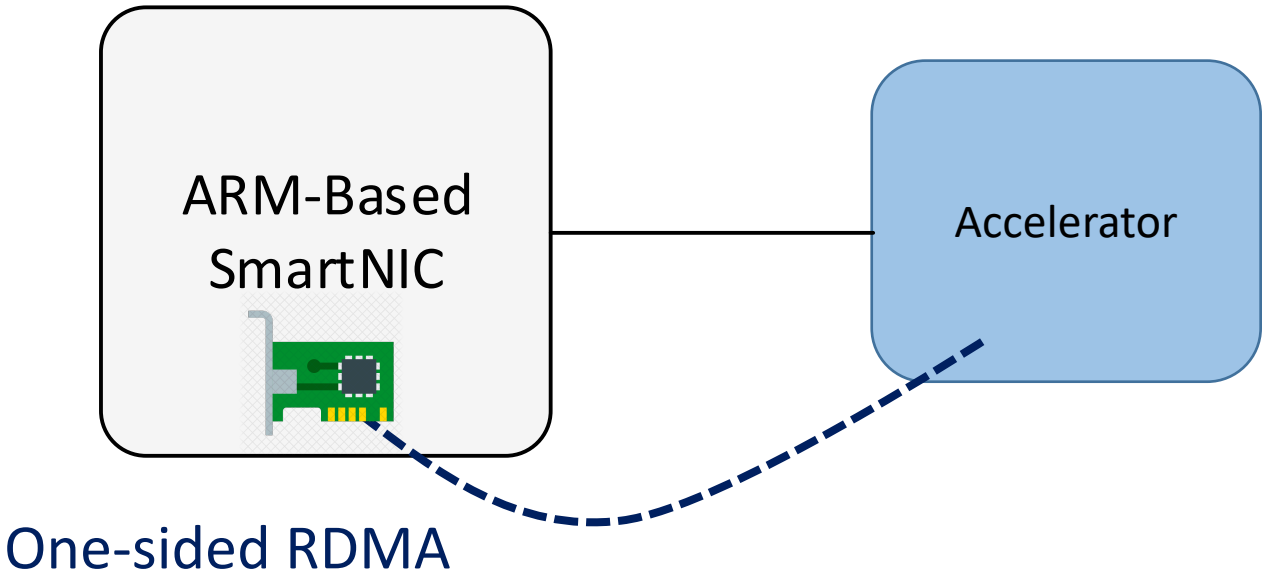
Why not using the accelerated RMDA engine integrated in the SmartNIC?



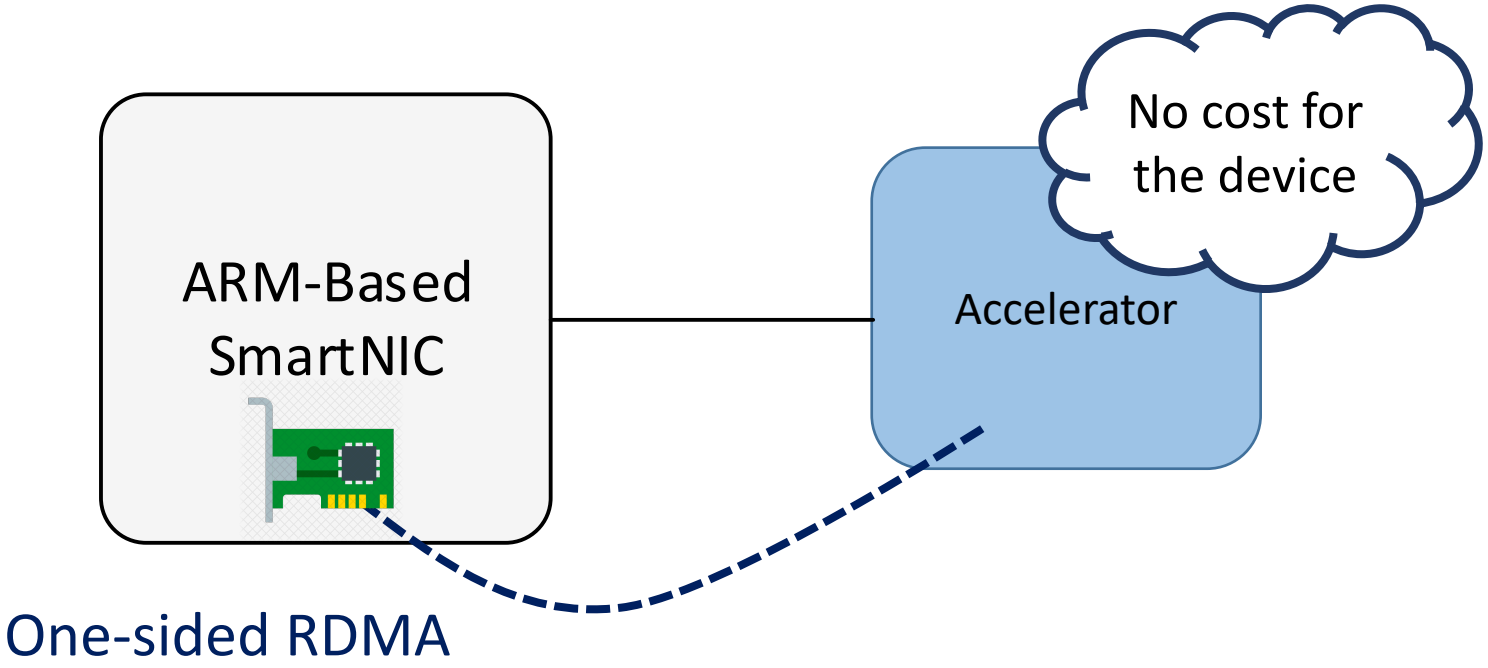
How can we do it?



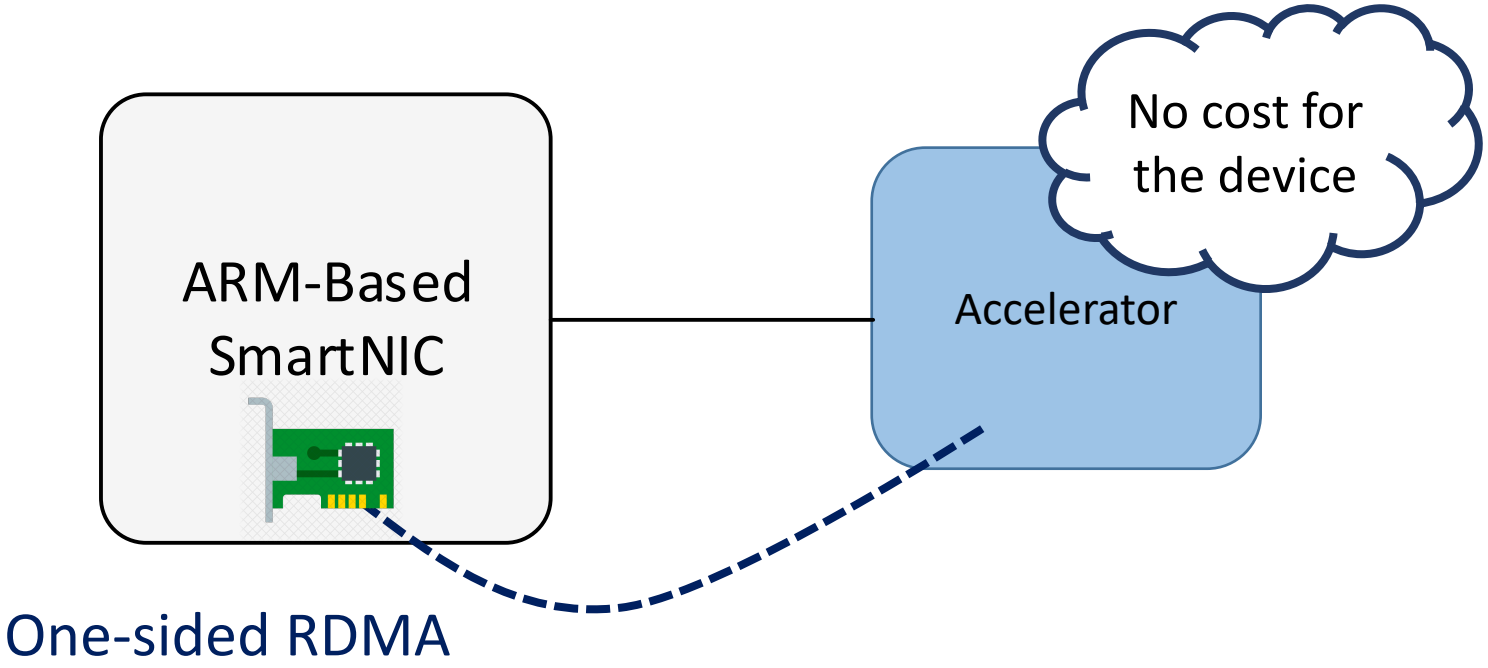
How can we do it?



How can we do it?

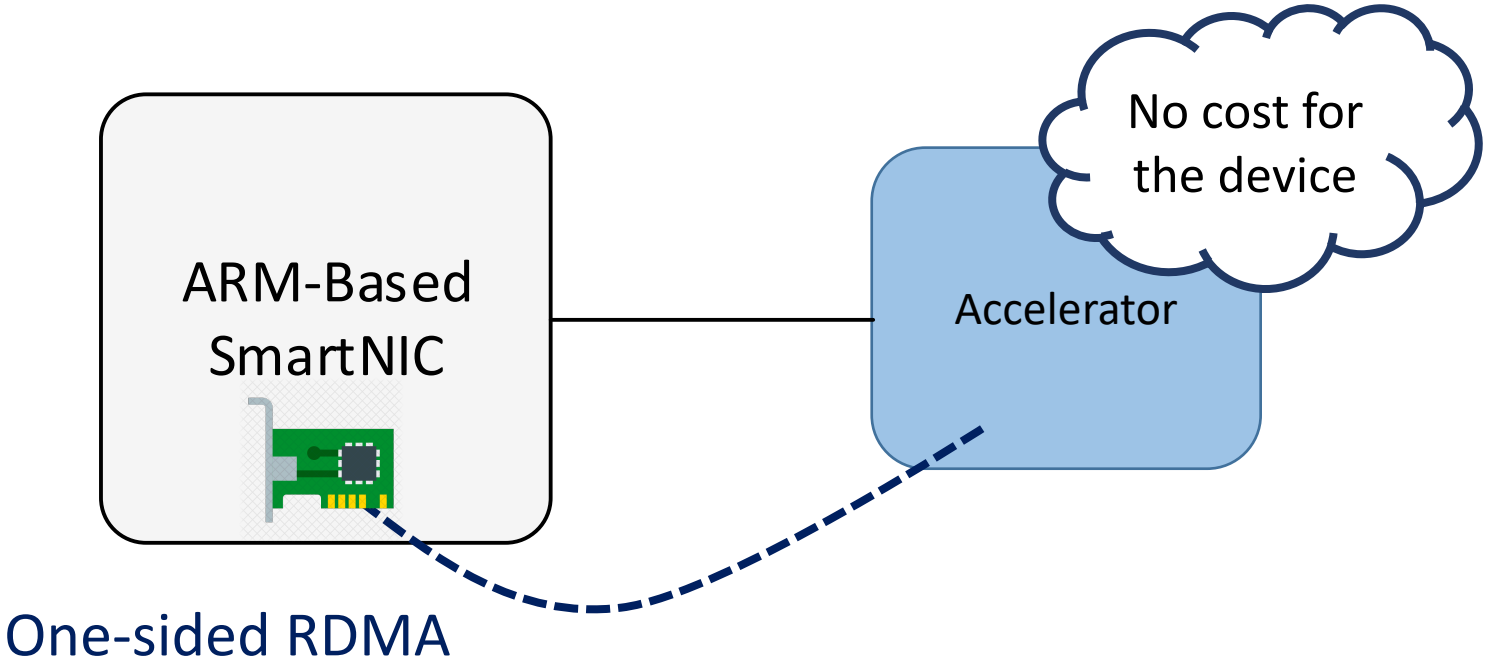


How can we do it?



Portable

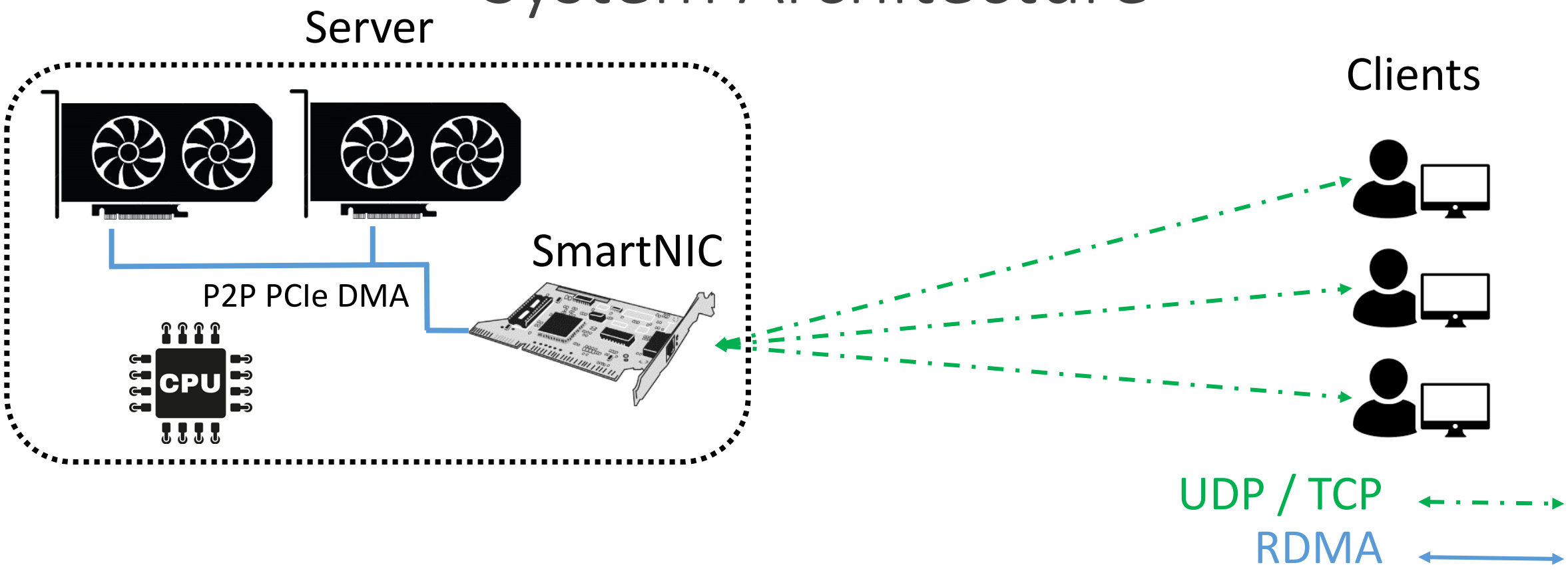
How can we do it?



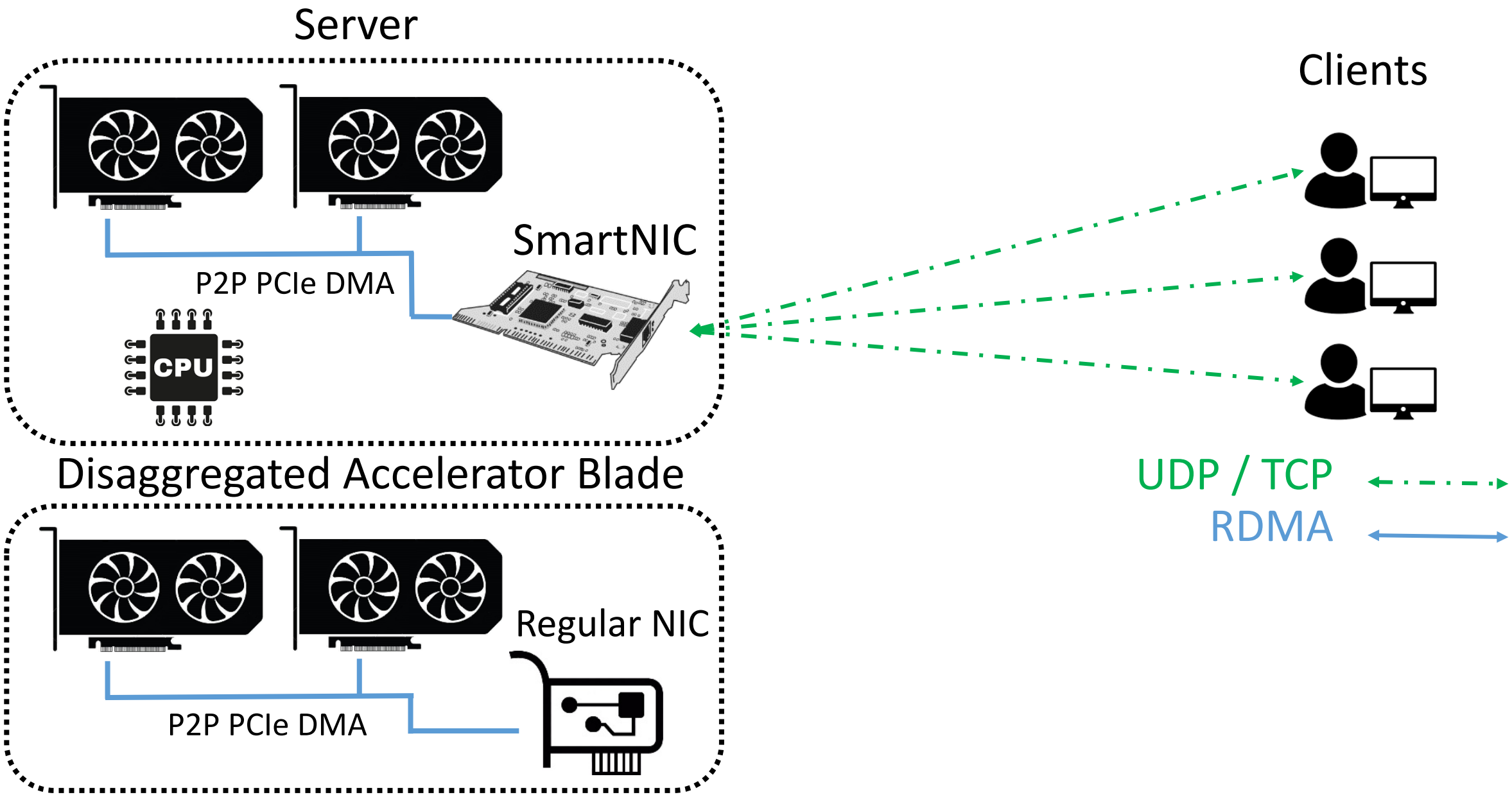
Portable

Data transfer and accelerator managment using RDMA outperforms existing transfer mechanisms in GPUs.
(see the paper)

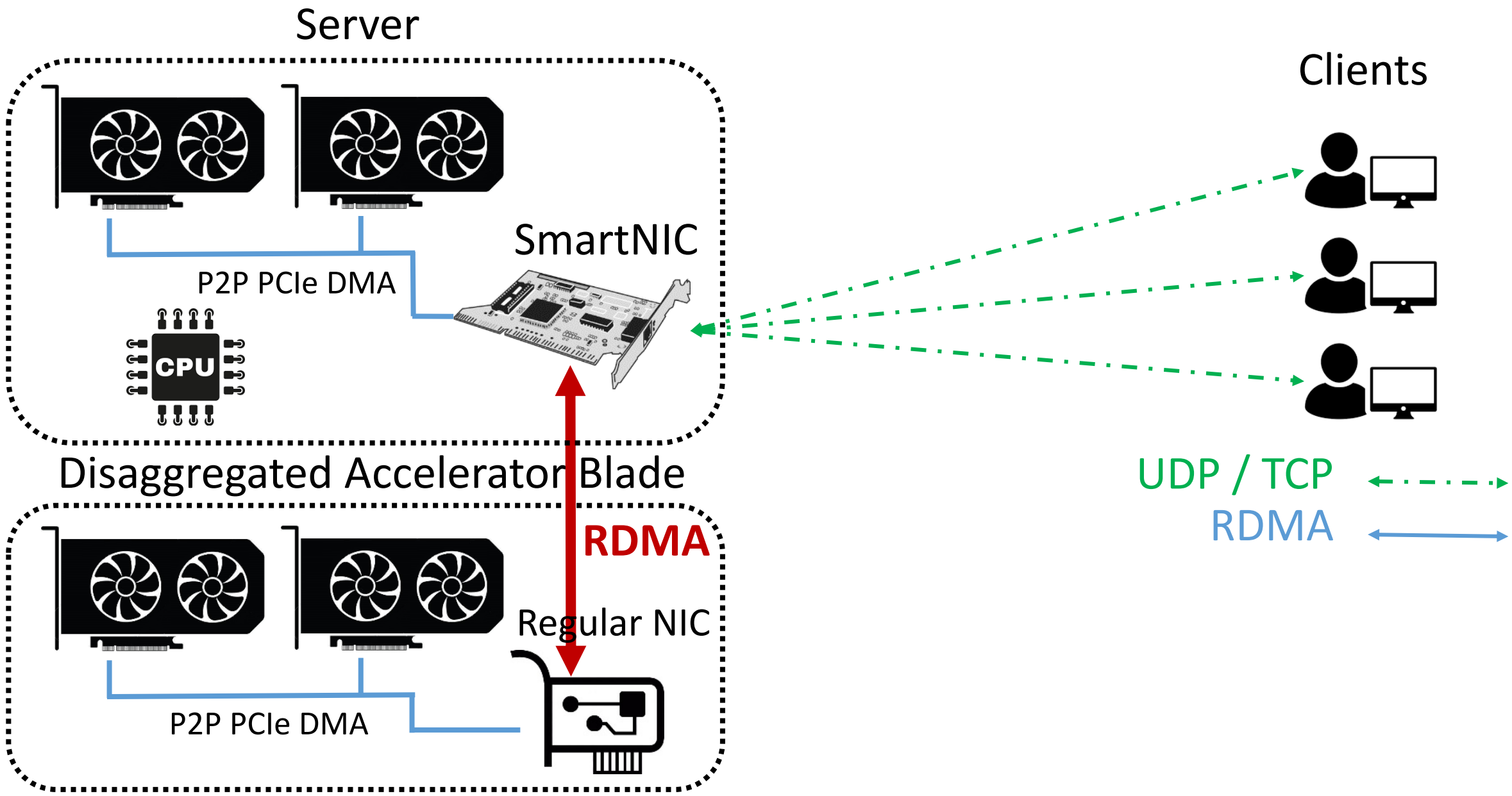
System Architecture



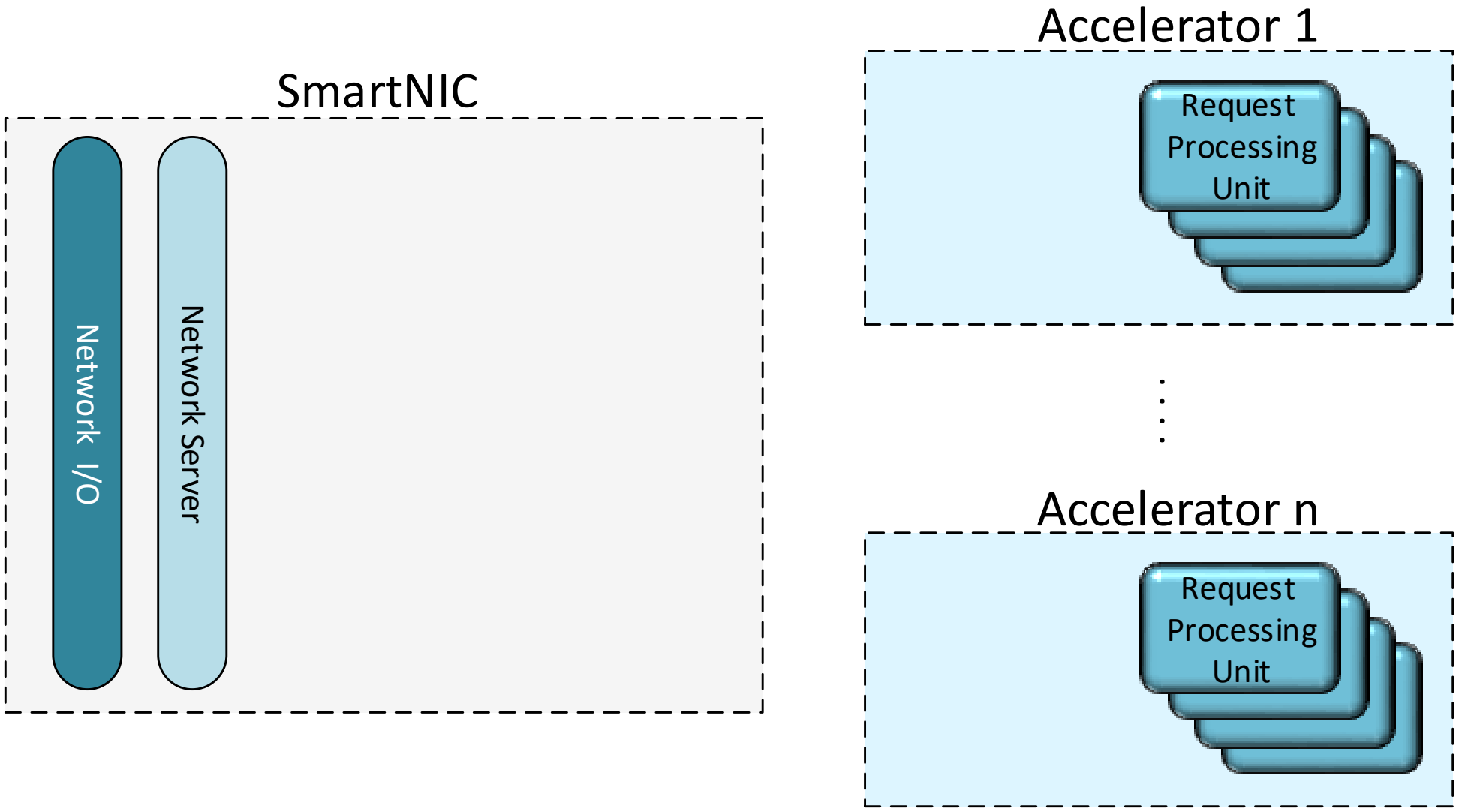
Seamless Scaling to Remote Accelerators



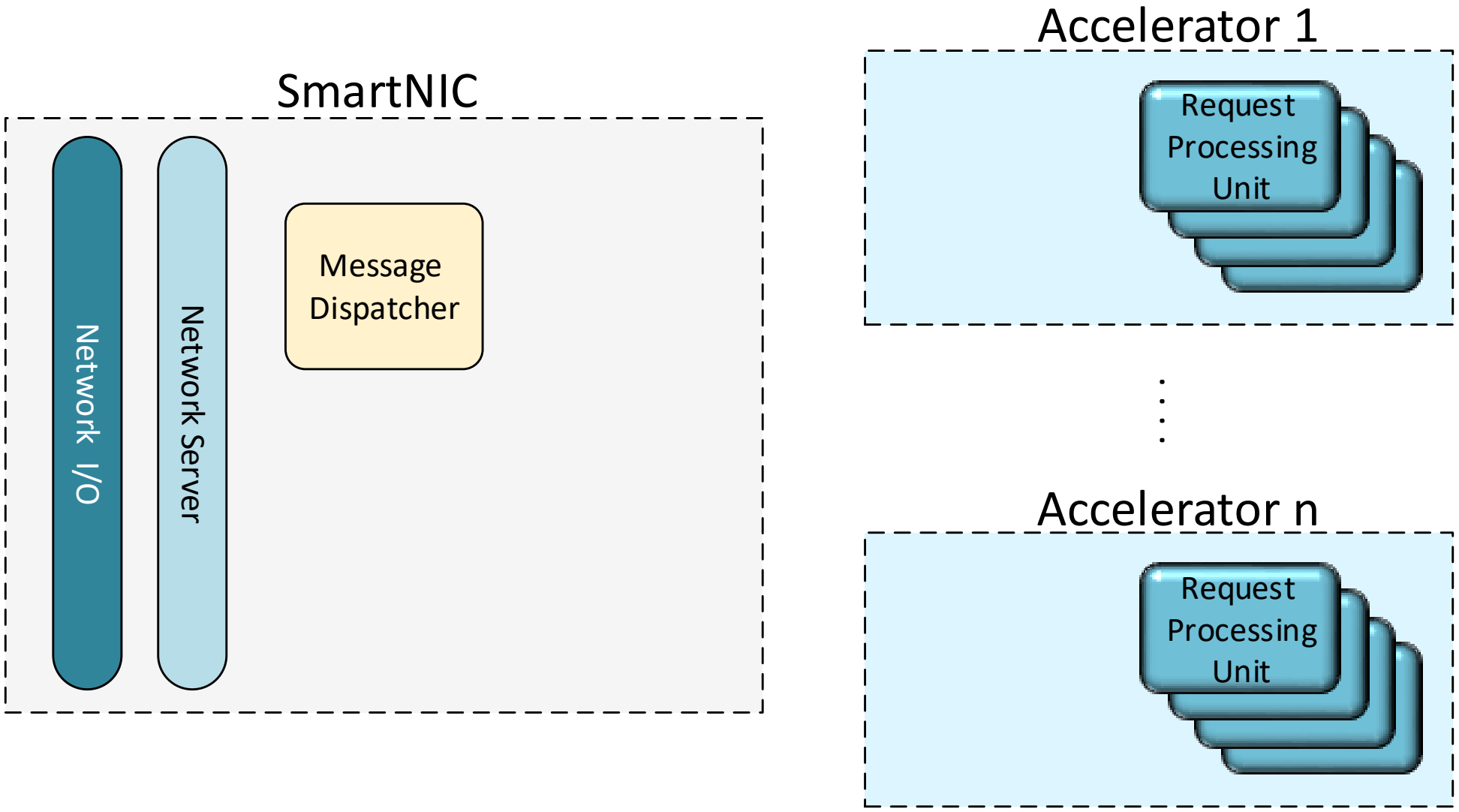
Seamless Scaling to Remote Accelerators



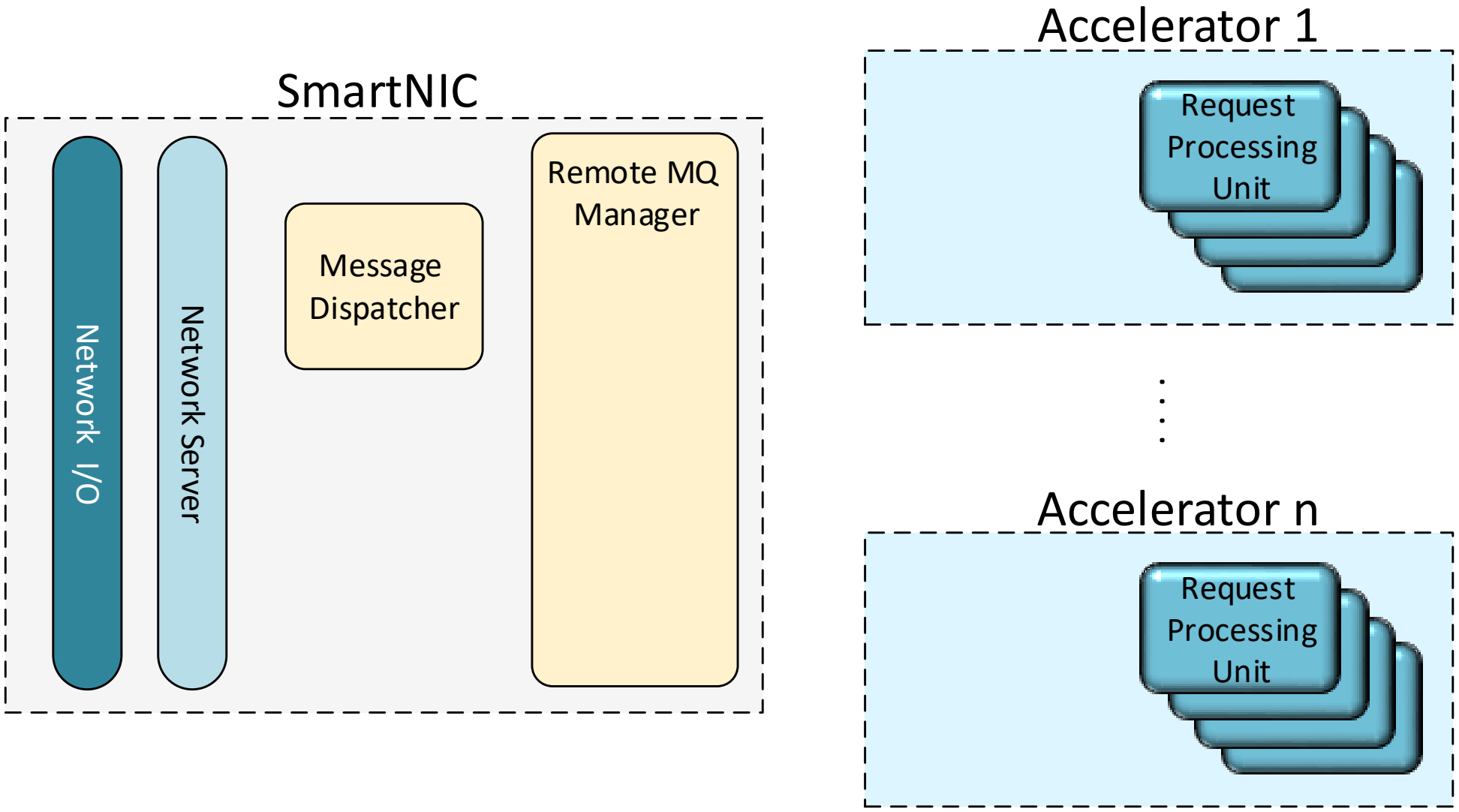
Lynx Components



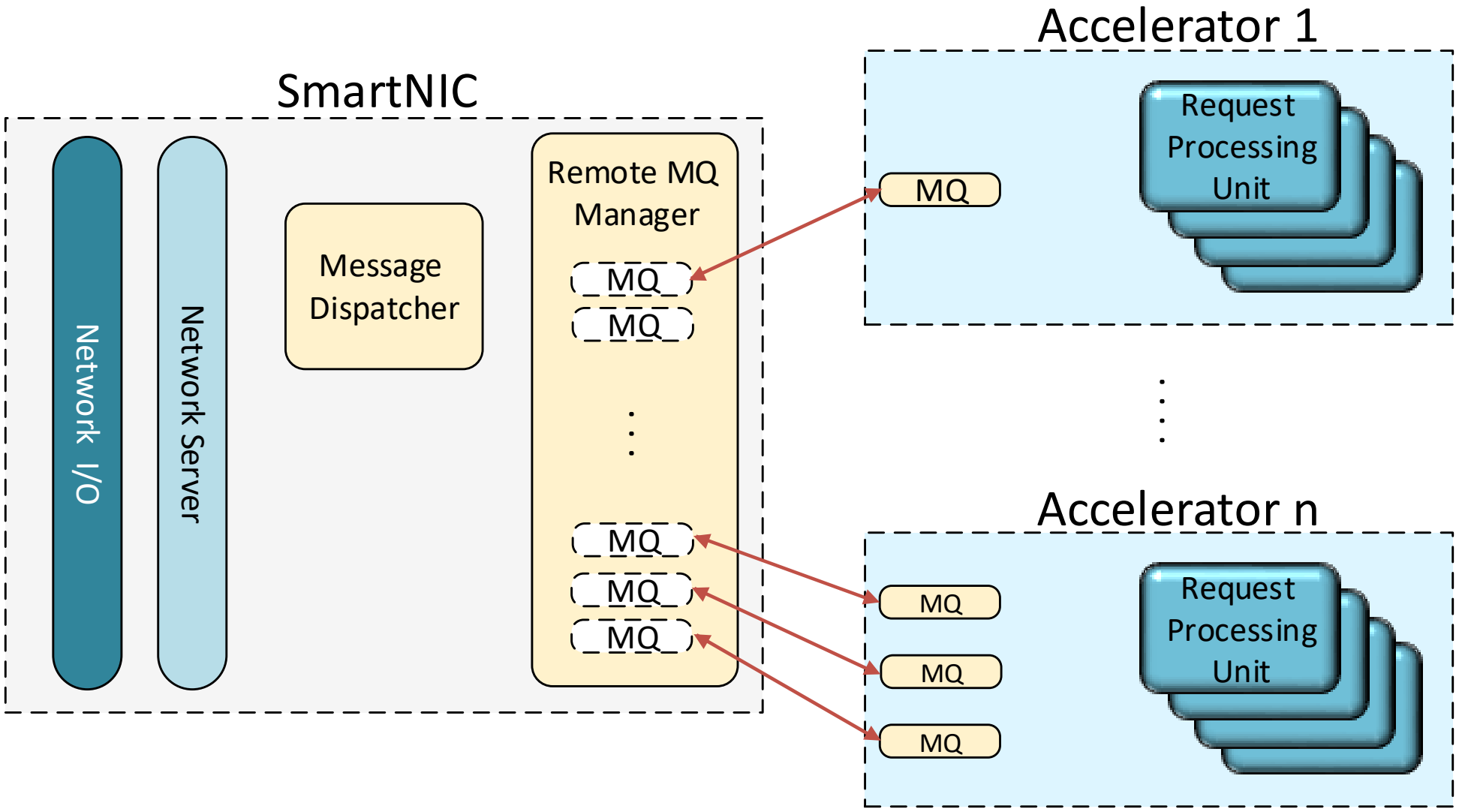
Lynx Components



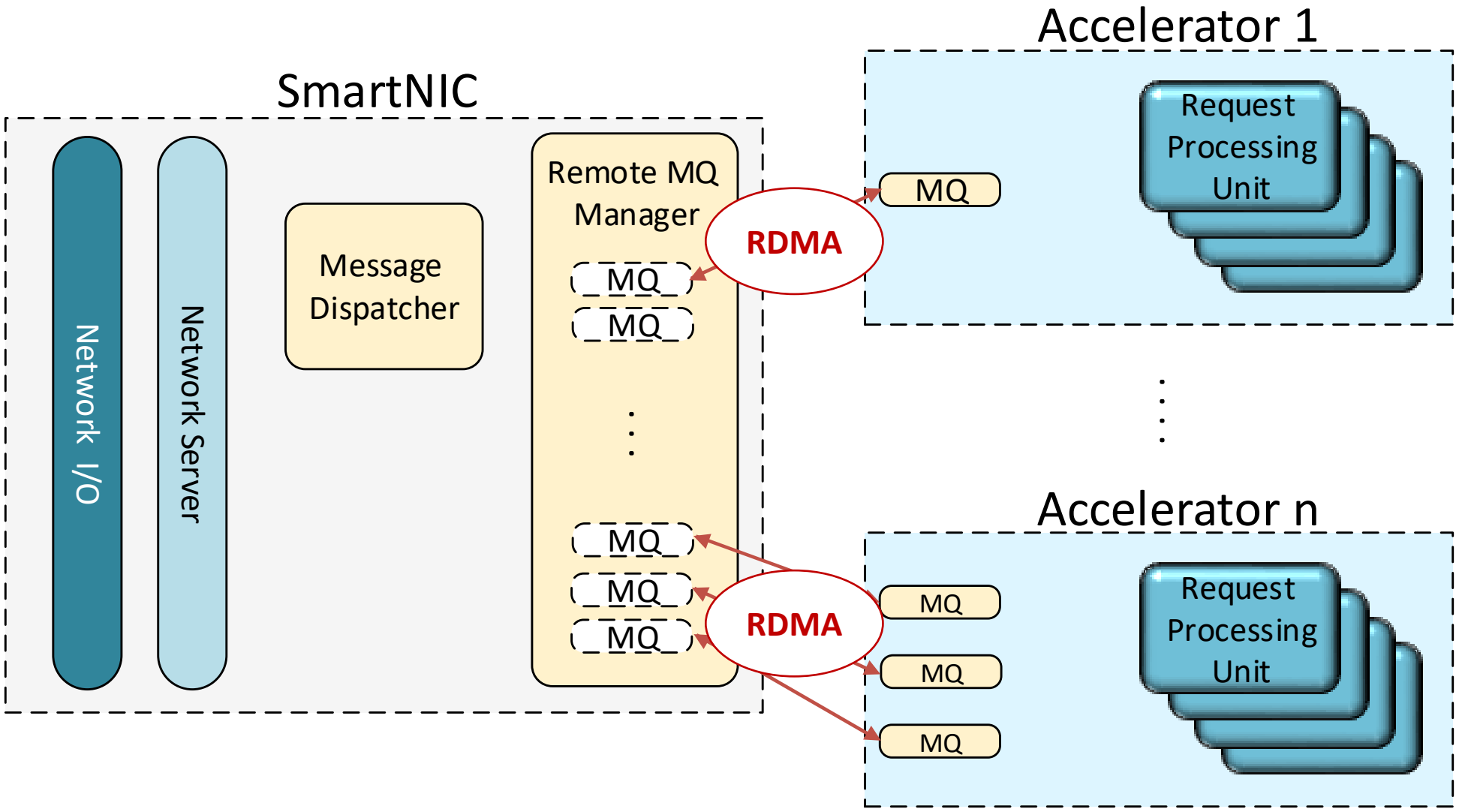
Lynx Components



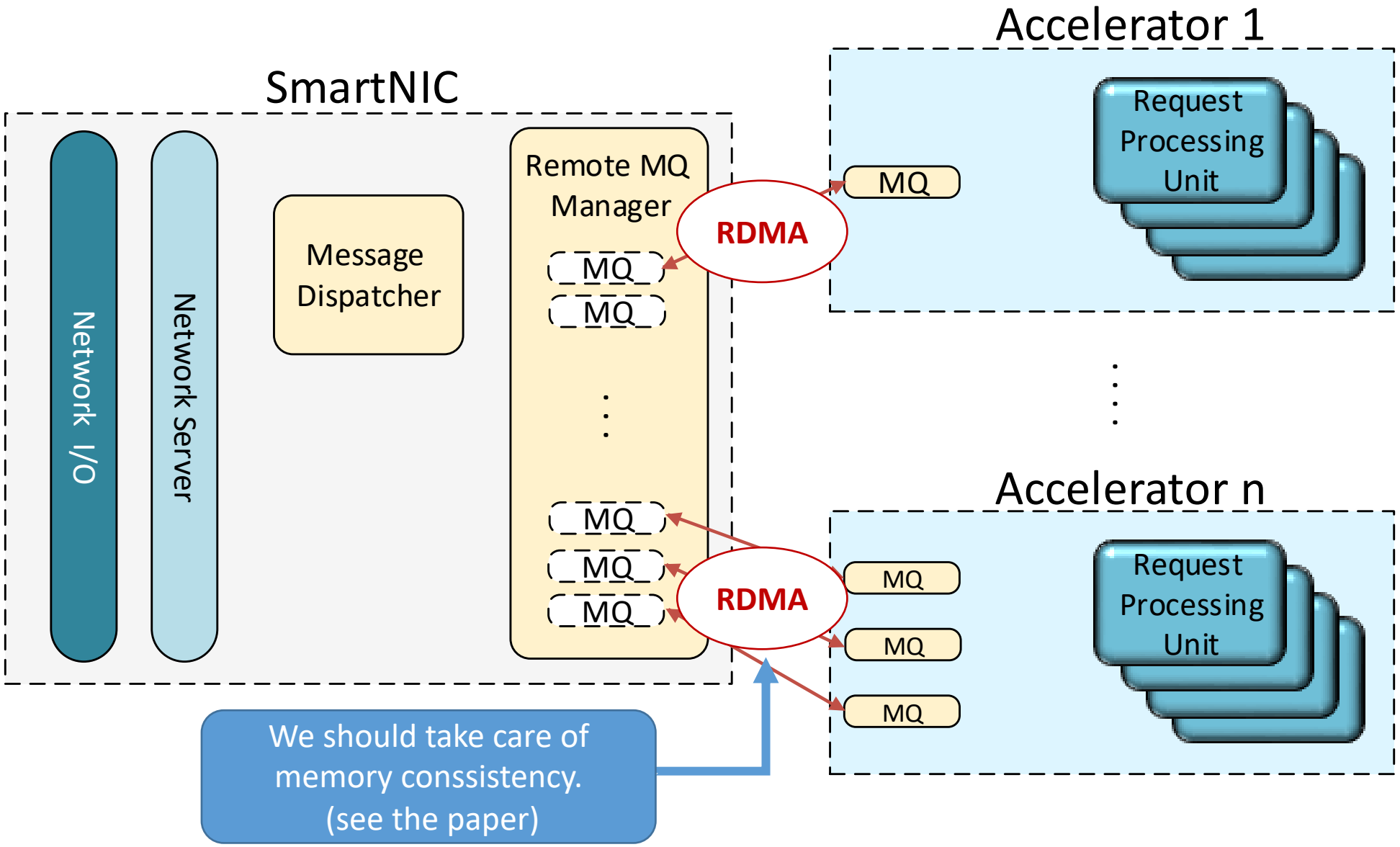
Lynx Components



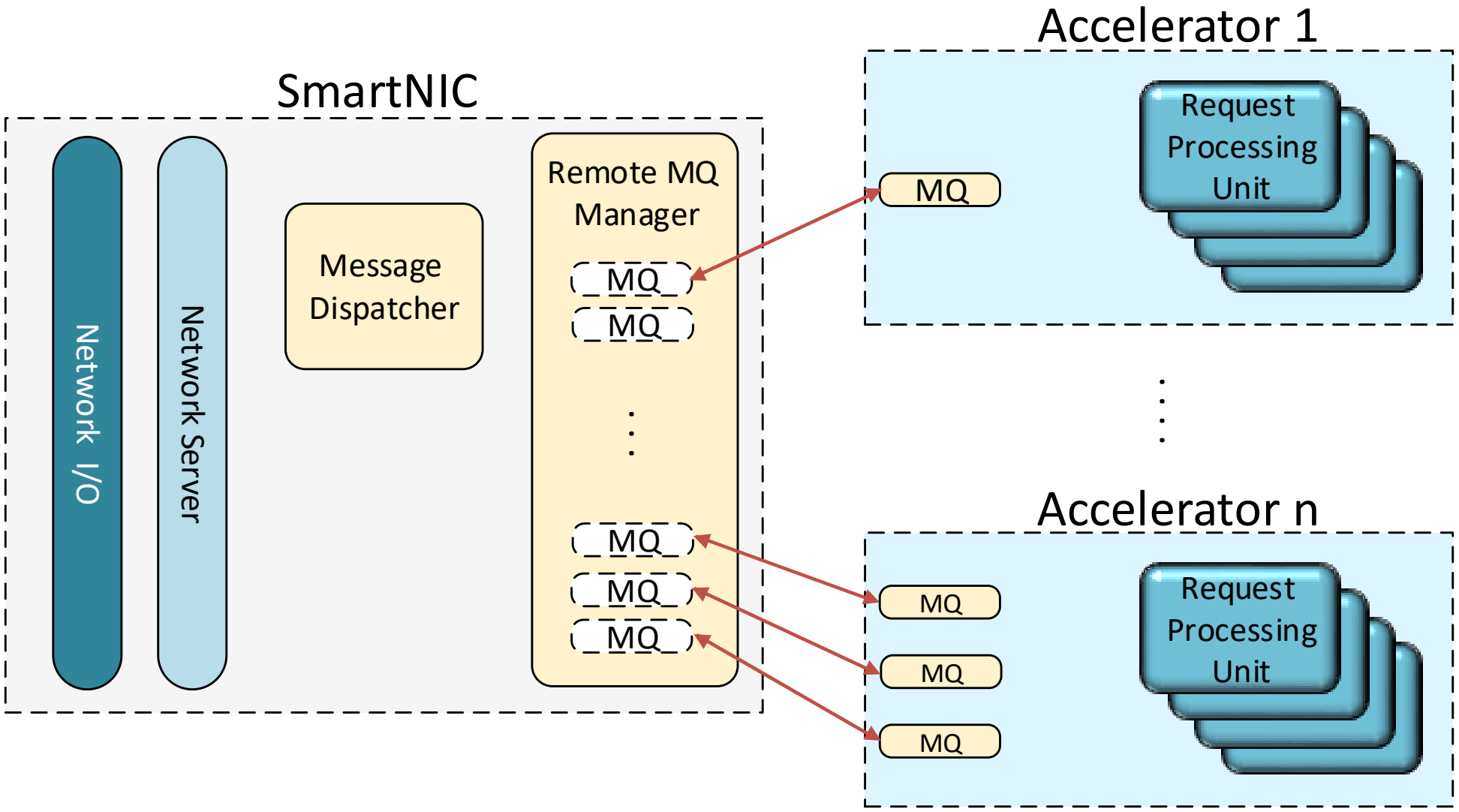
Lynx Components



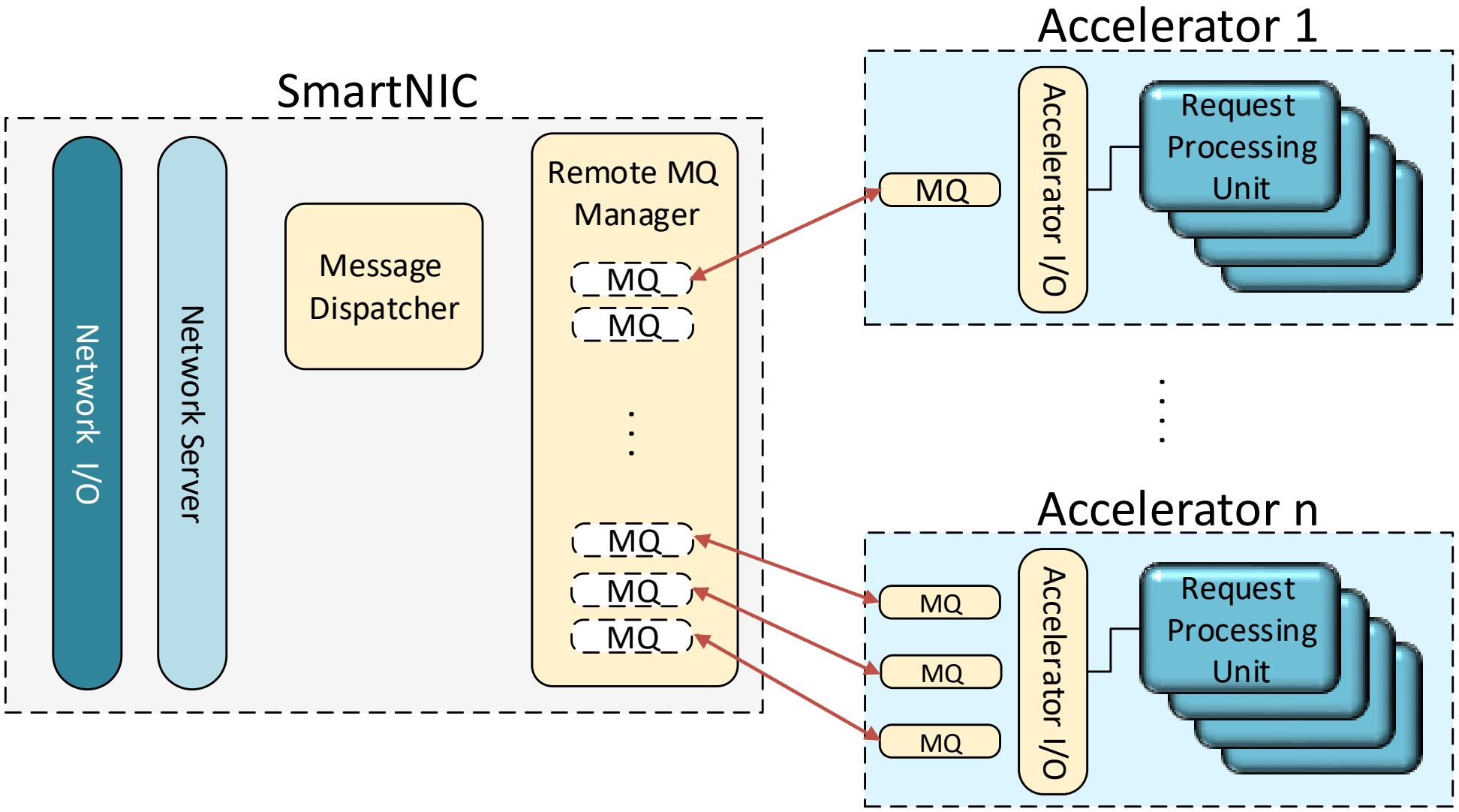
Lynx Components



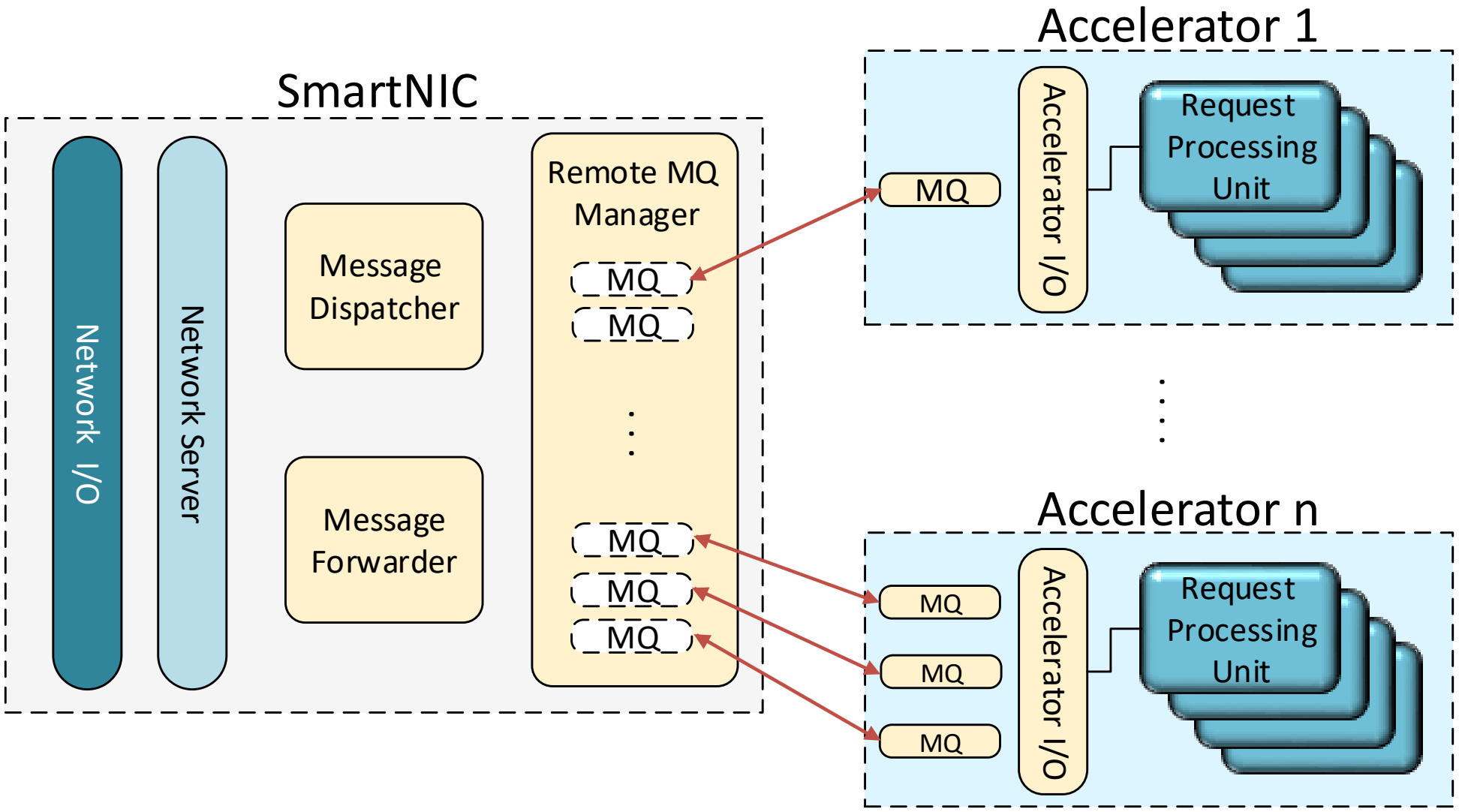
Lynx Components



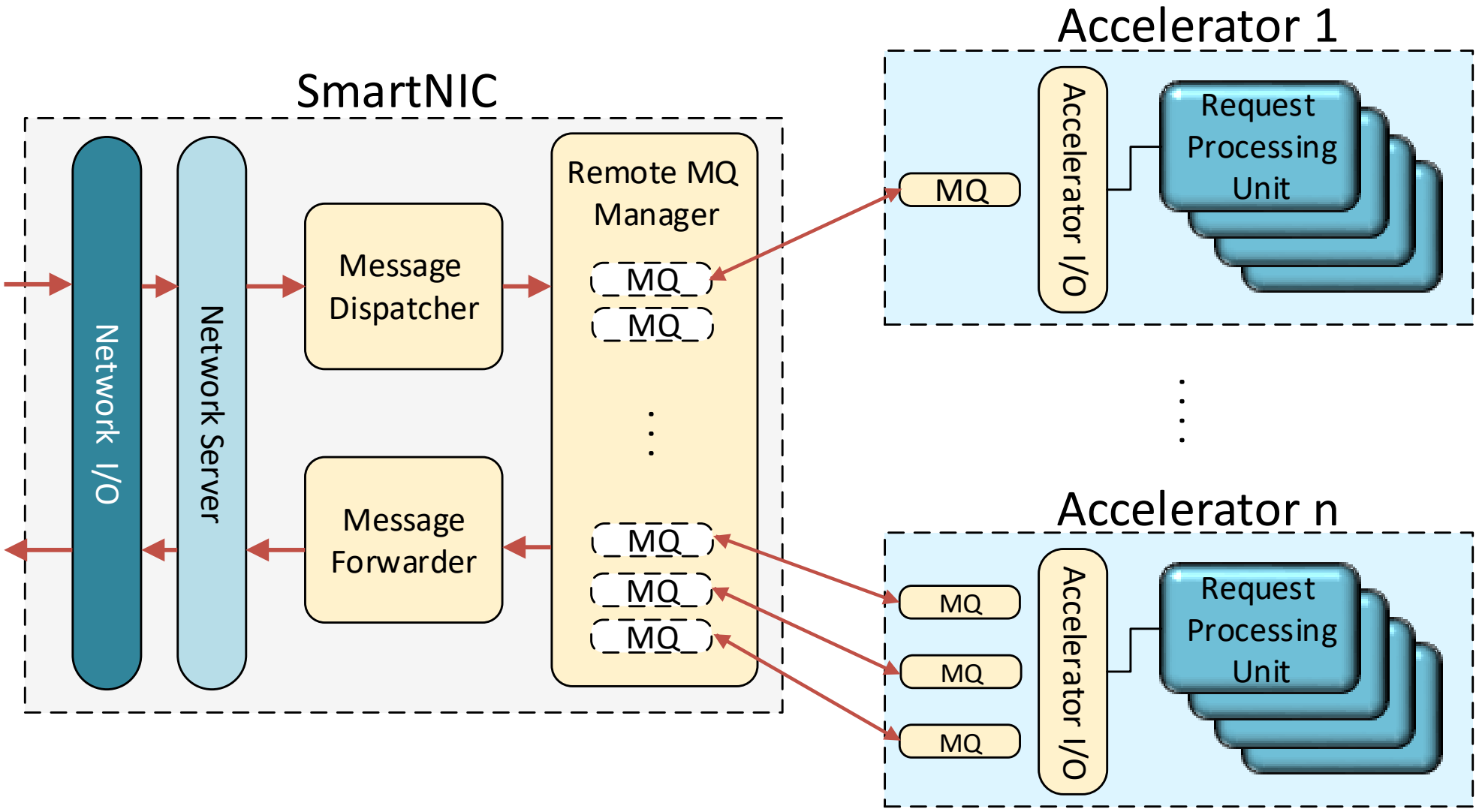
Lynx Components



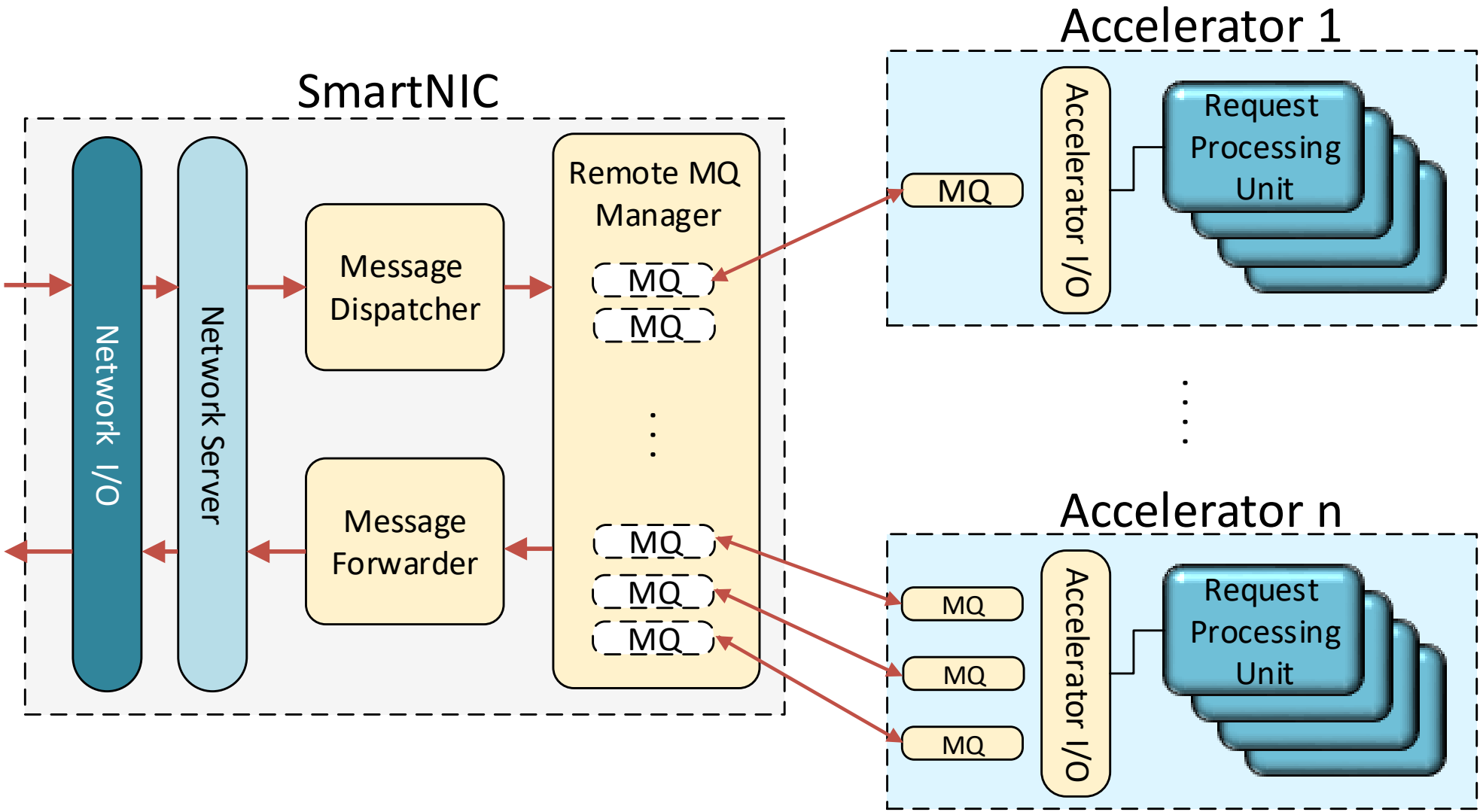
Lynx Components



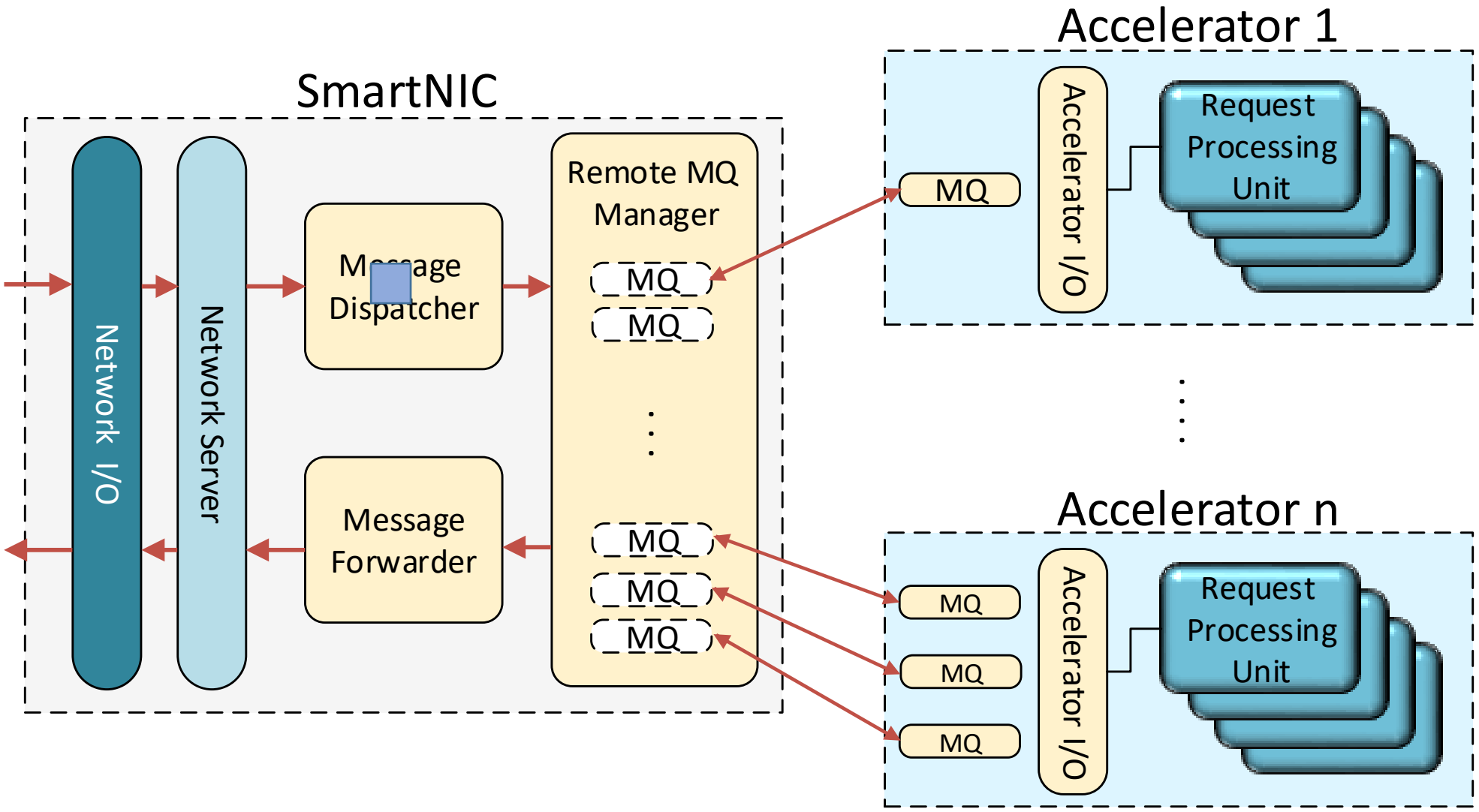
Lynx Components



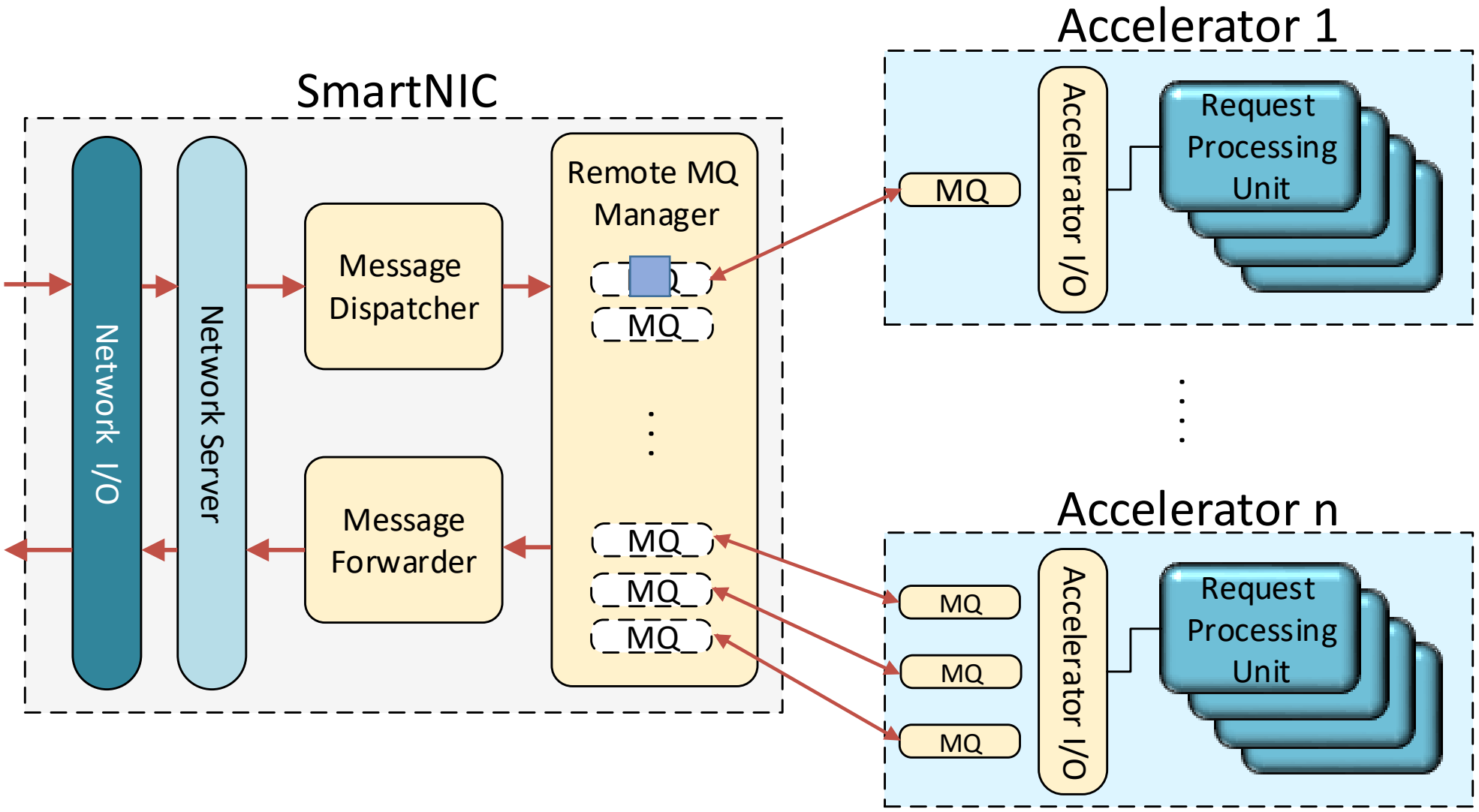
Lynx Components



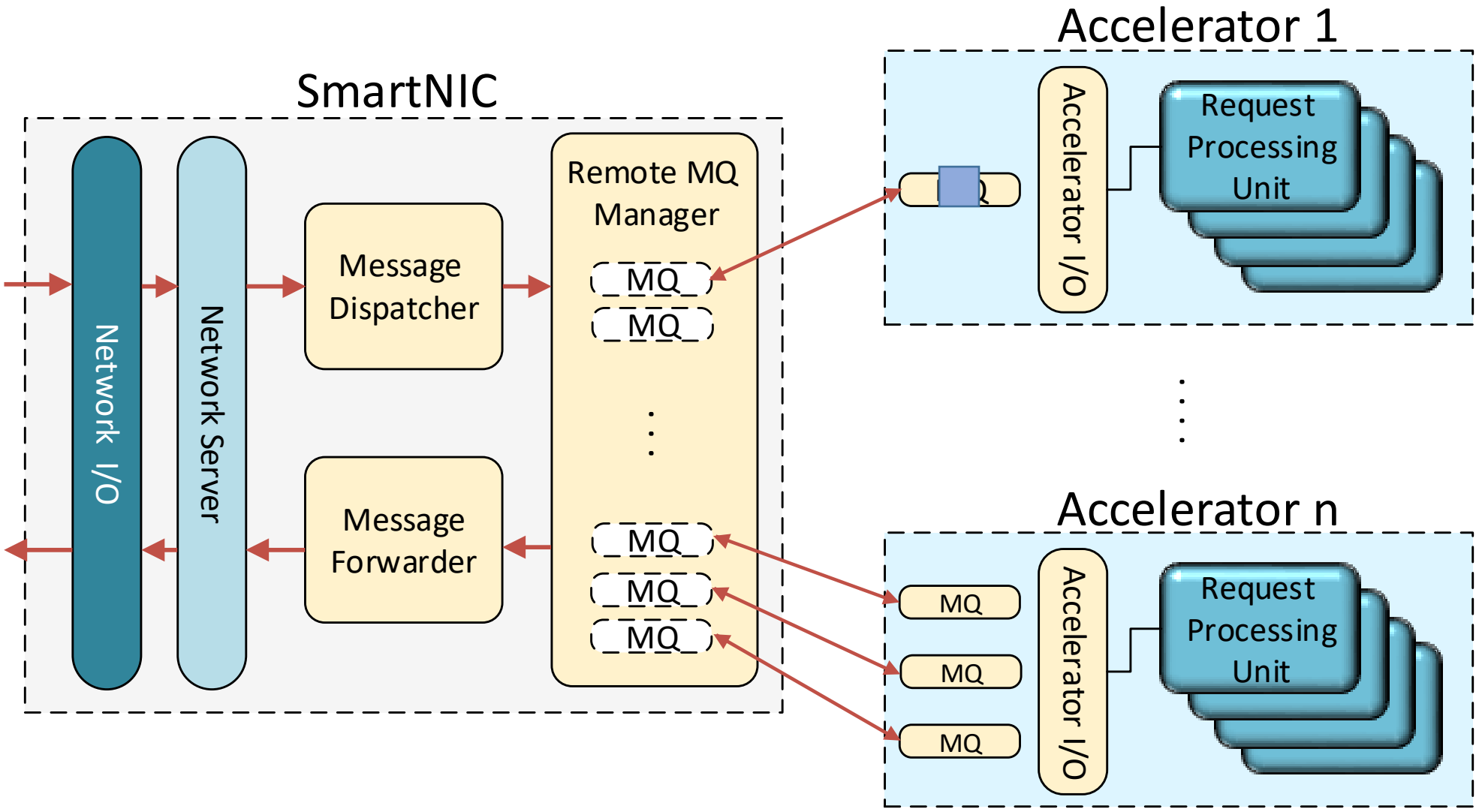
Lynx Components



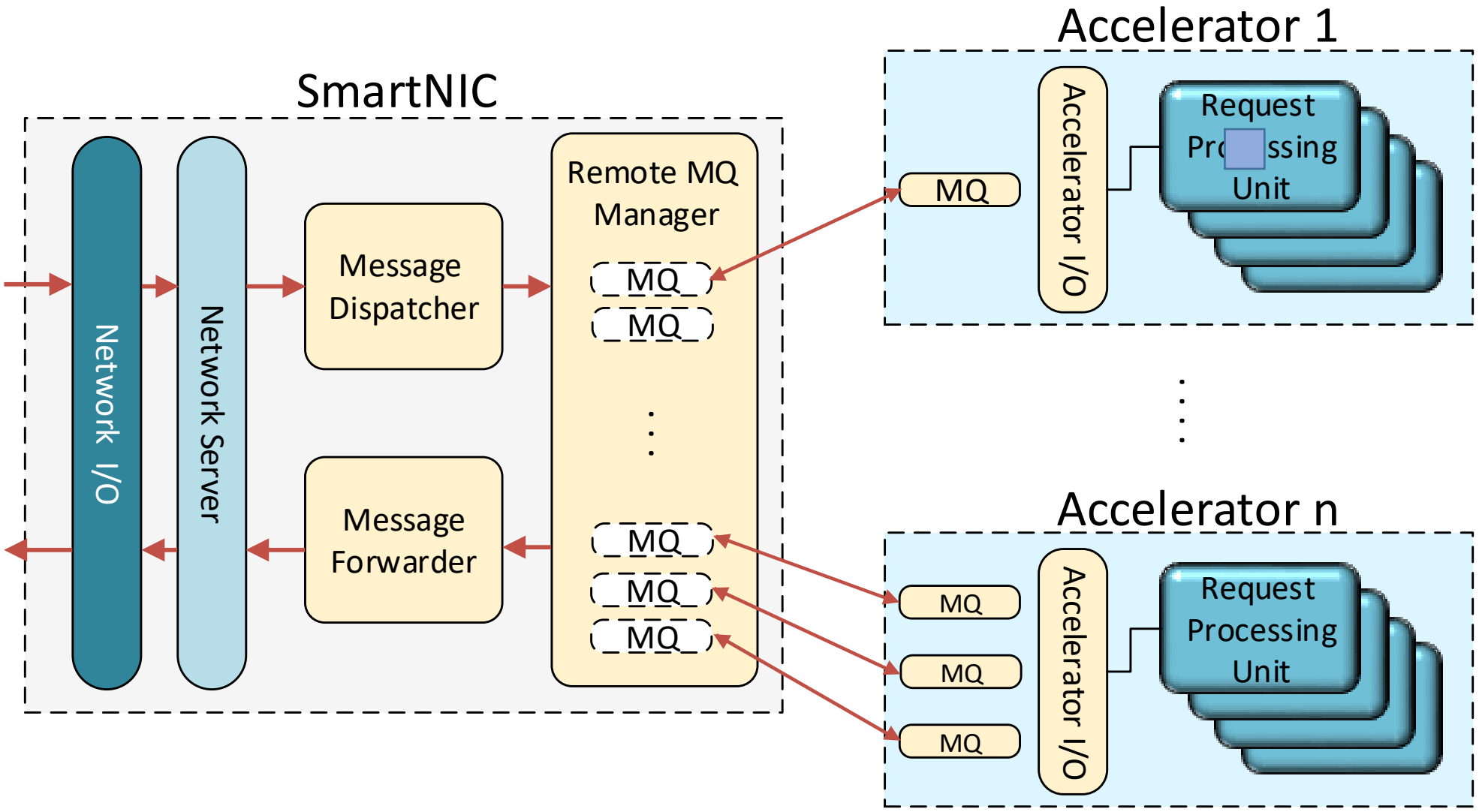
Lynx Components



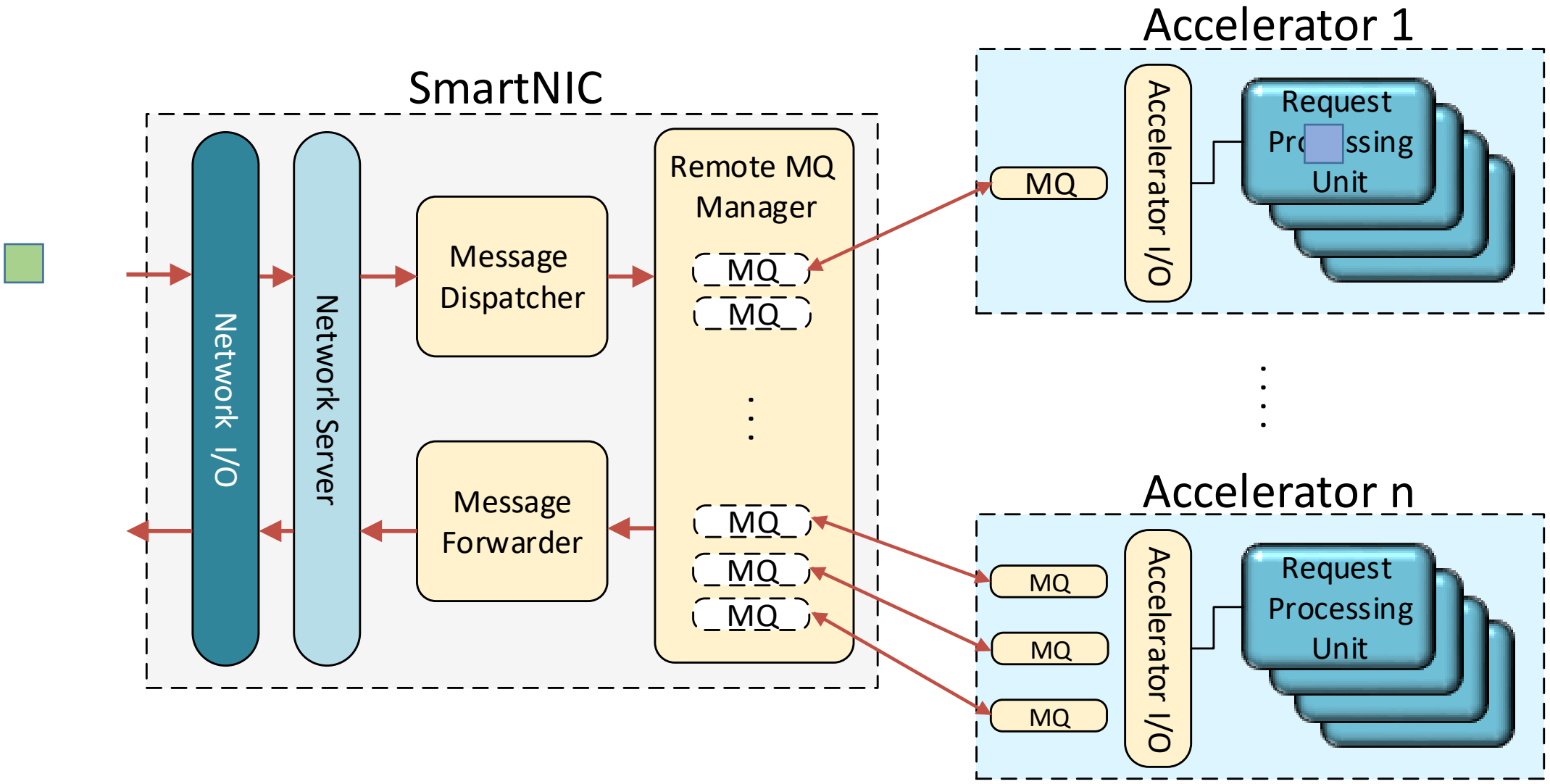
Lynx Components



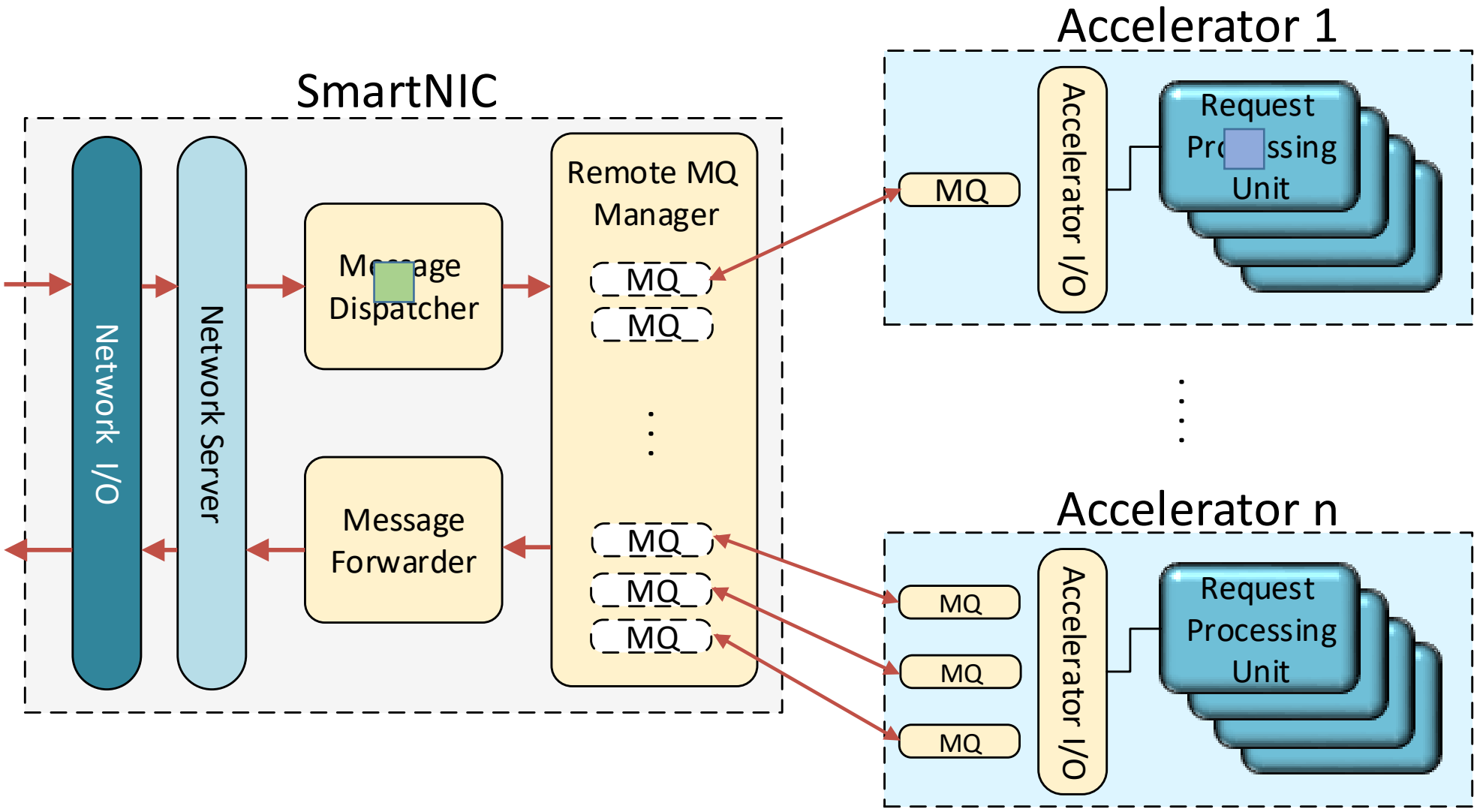
Lynx Components



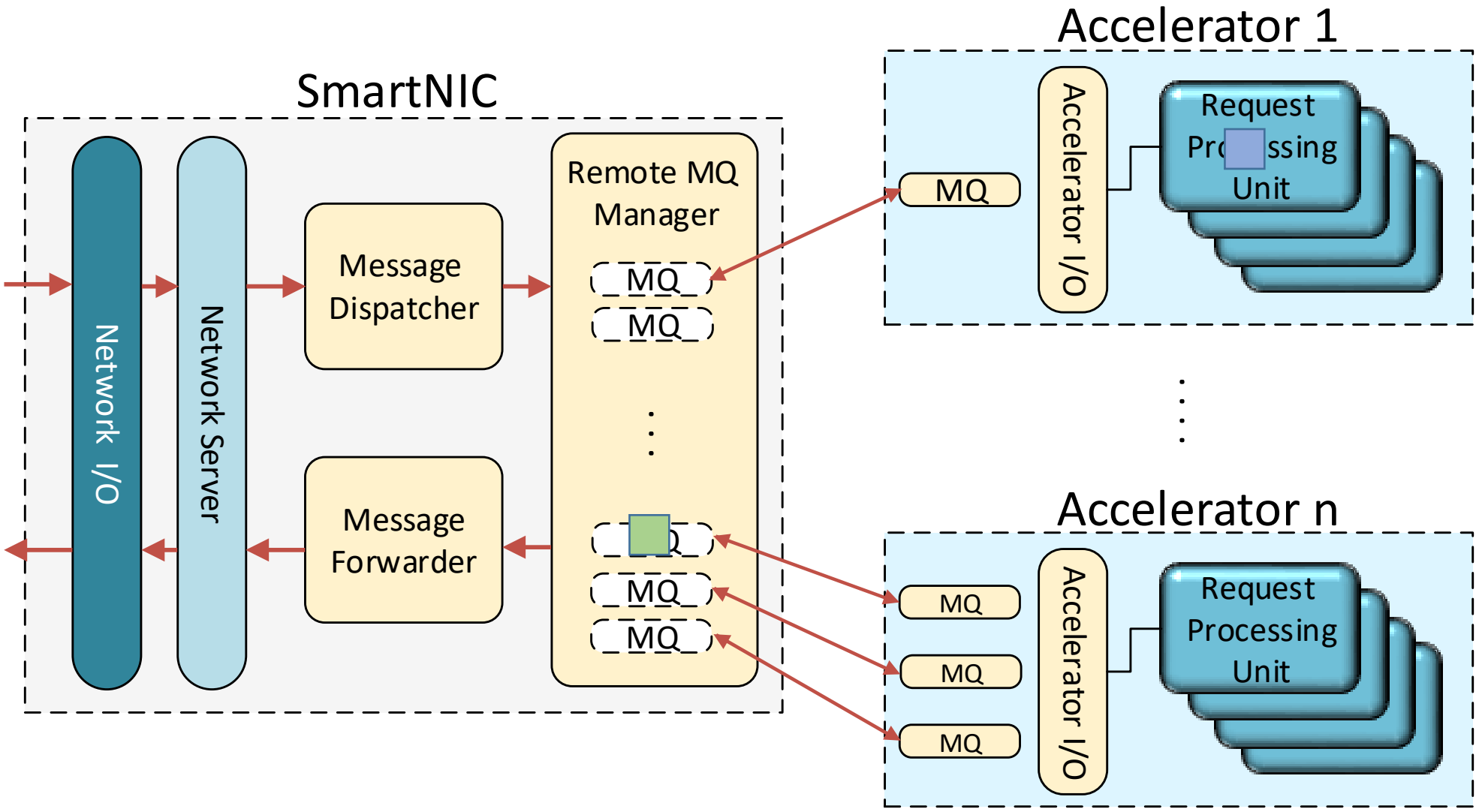
Lynx Components



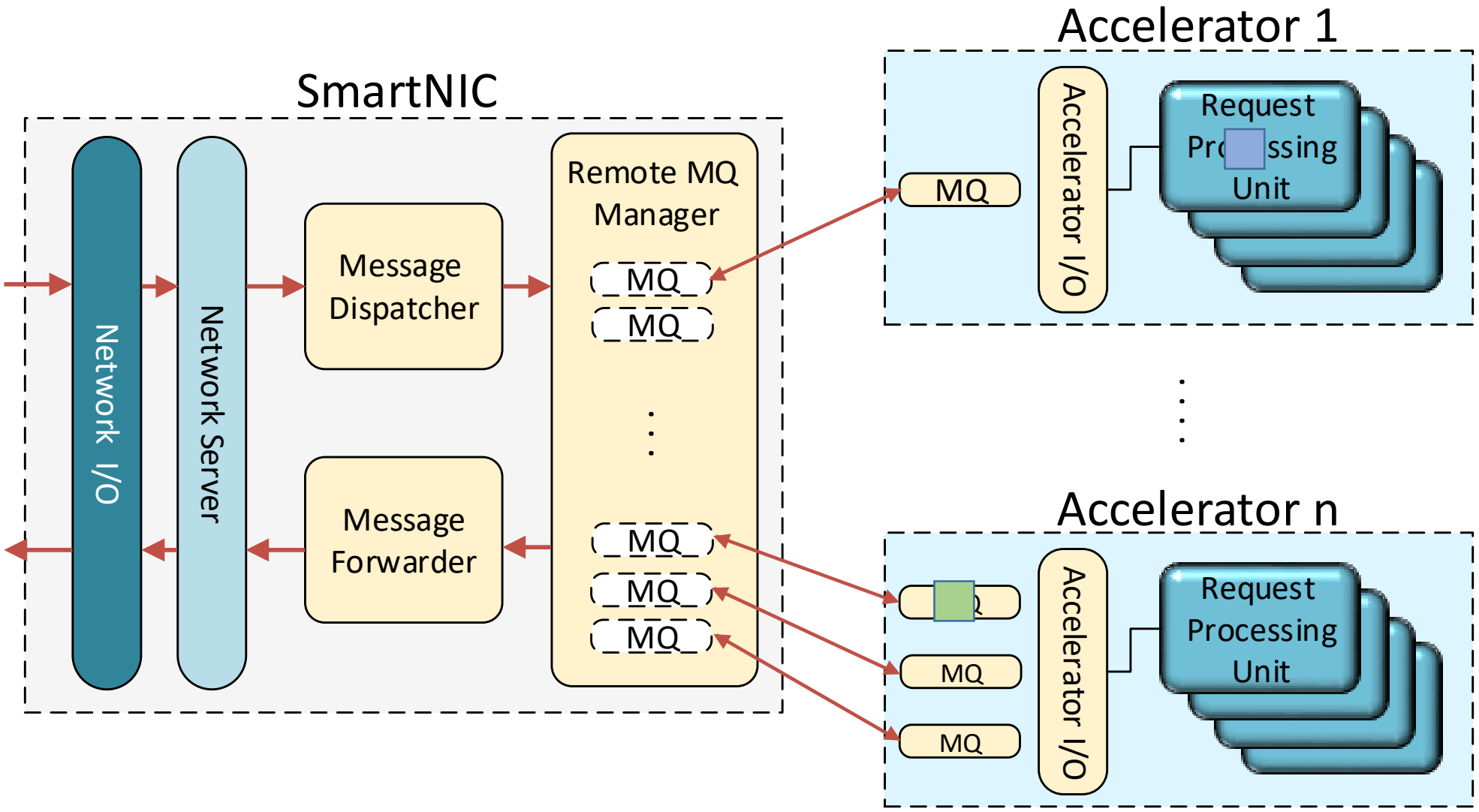
Lynx Components



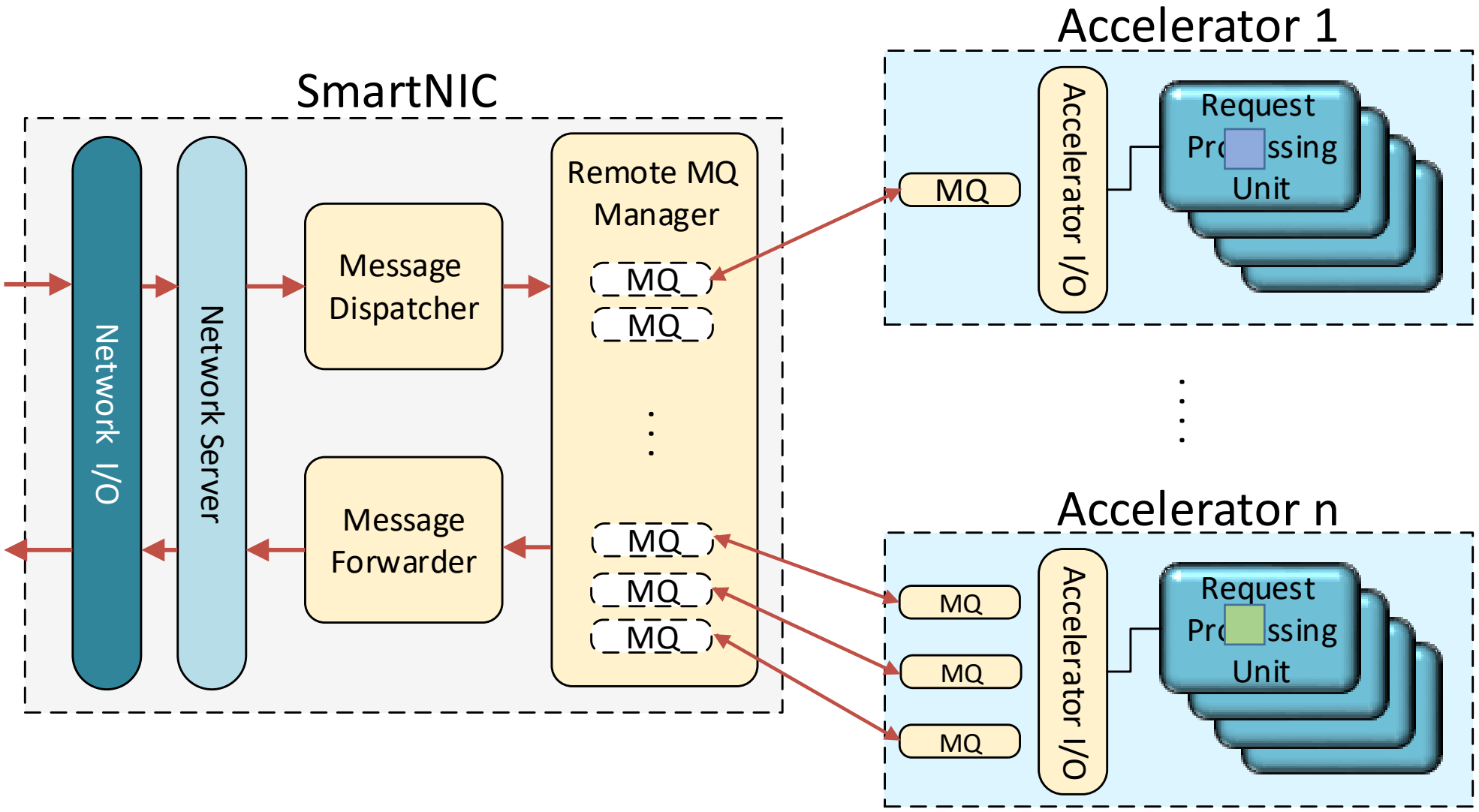
Lynx Components



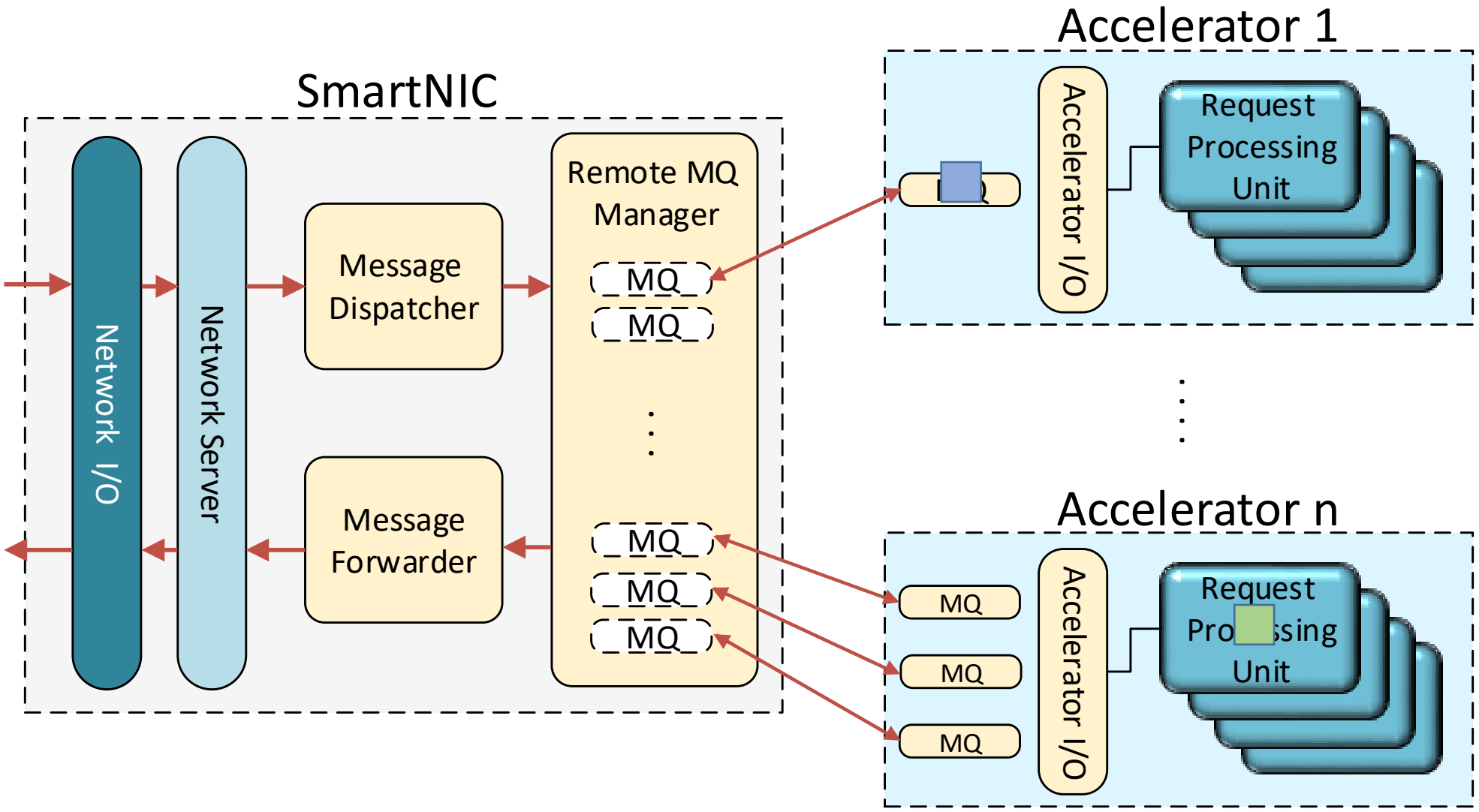
Lynx Components



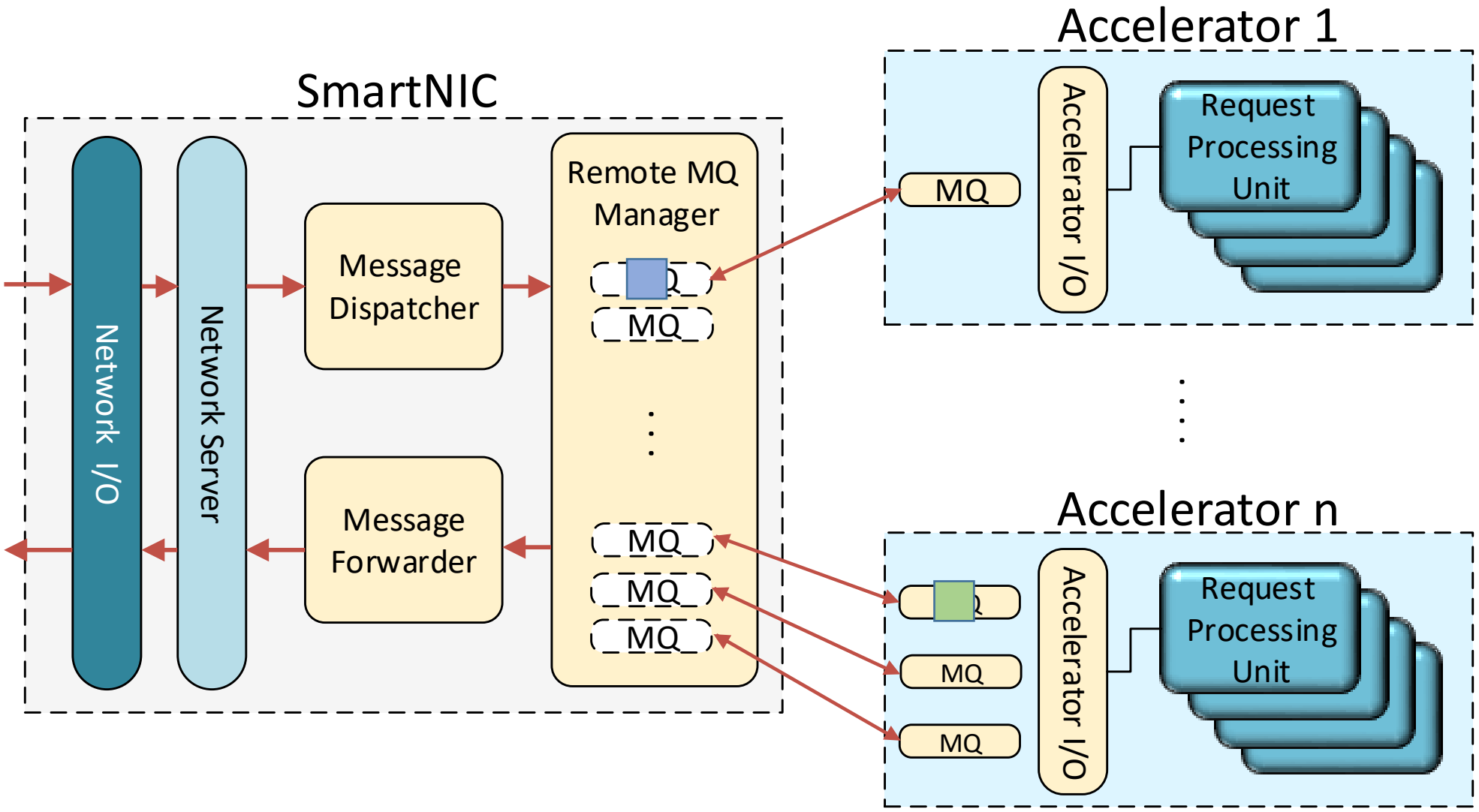
Lynx Components



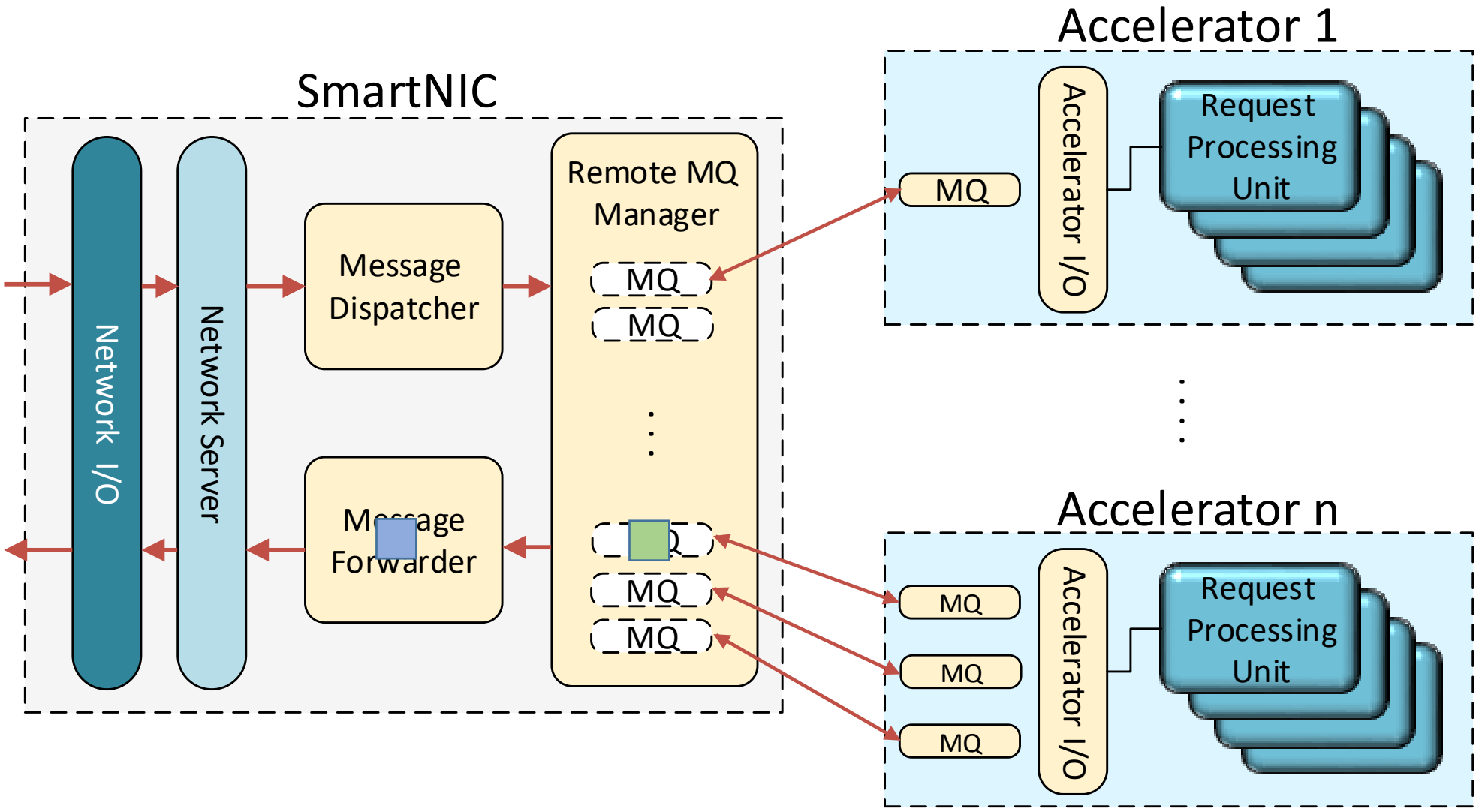
Lynx Components



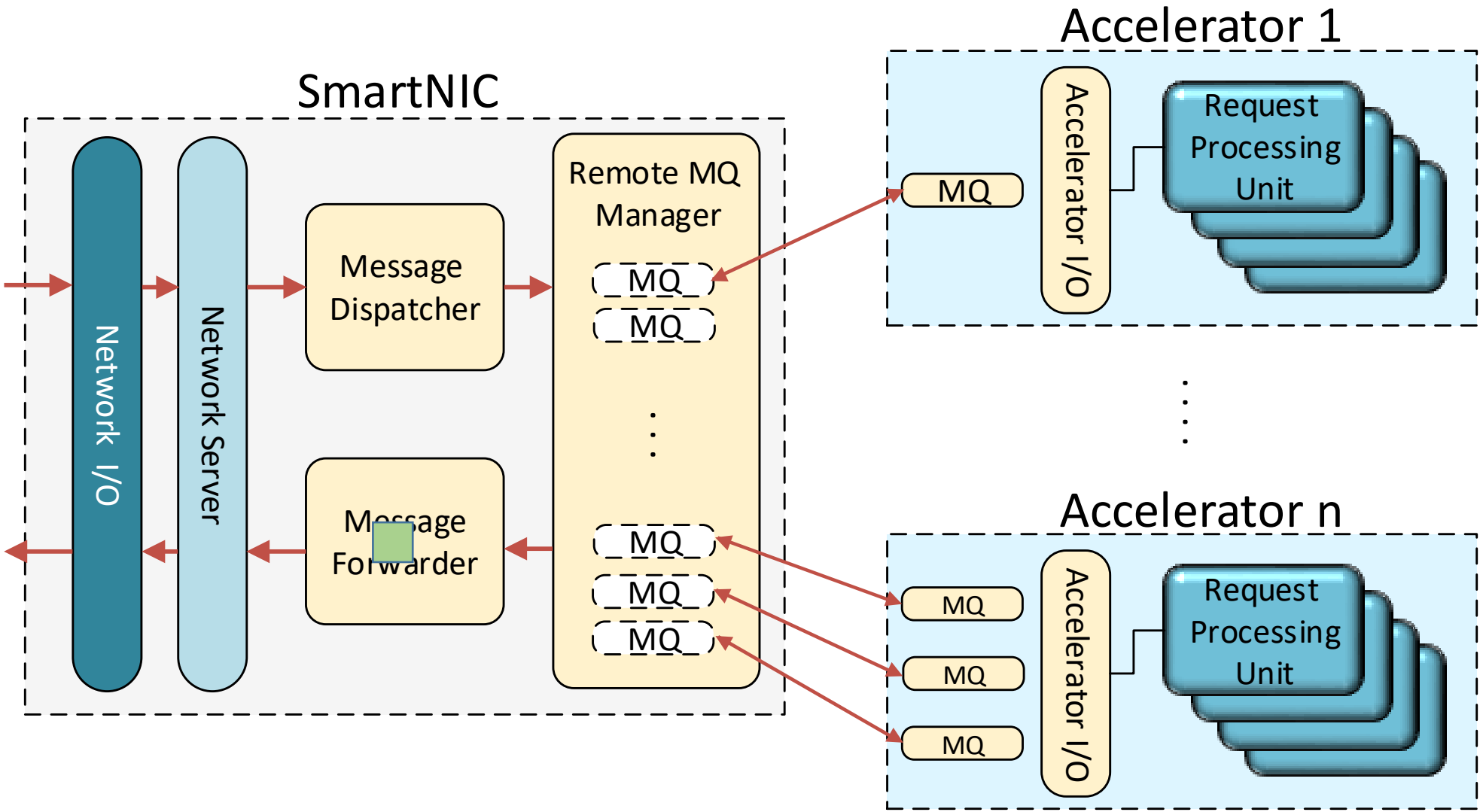
Lynx Components



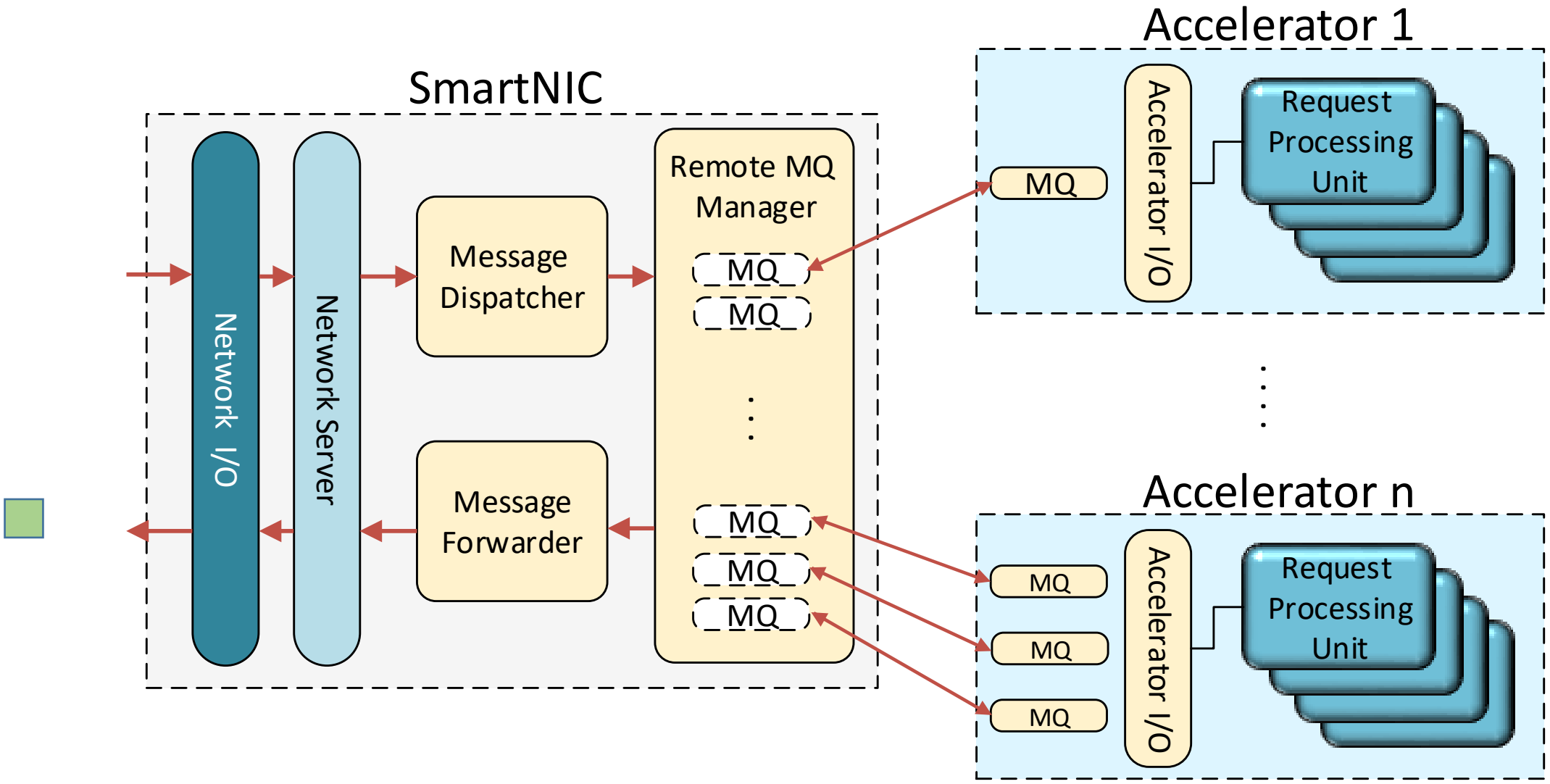
Lynx Components



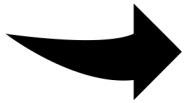
Lynx Components



Lynx Components



Agenda



- Motivation and Background
- Design
- Evaluation
- Conclusions

Implementation

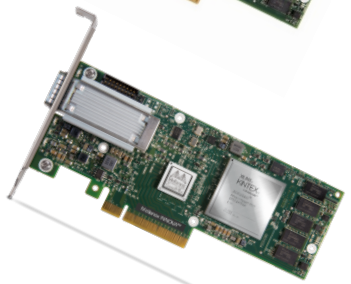
SmartNICs

- ARM-based (Bluefield)
- FPGA-based (Innova)

Implementation

SmartNICs

- ARM-based (Bluefield)
- FPGA-based (Innova)



Accelerators

- NVIDIA GPU

- Intel Visual Computer Accelerator – VCA

Experiments

- Micro benchmarks
  - Throughput, Latency, Isolation
- Face Verification Server
  - Accelerator-initiated I/O, multiple transport protocols
- DNN Model inference
  - CPU-less low-latency service, scalability
- Secure Server Computing inside SGX enclave
  - Portability: using Intel VCA

Experiments

- ❑ Micro benchmarks
  - ❑ Throughput, Latency, Isolation

- ❑ Face Verification Server
  - ❑ Accelerator-initiated I/O, multiple transport protocols

- ✓ DNN Model inference
  - ❑ CPU-less low-latency service, scalability

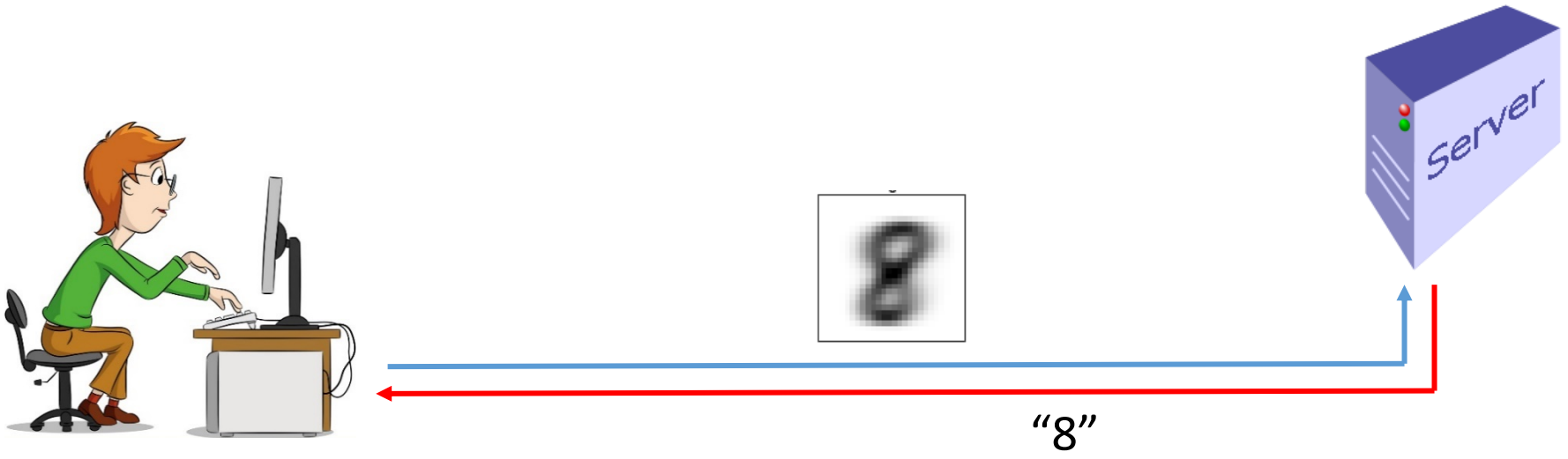
- ✓ Secure Server Computing inside SGX enclave
  - ❑ Portability: using Intel VCA

For more results, please read the paper.

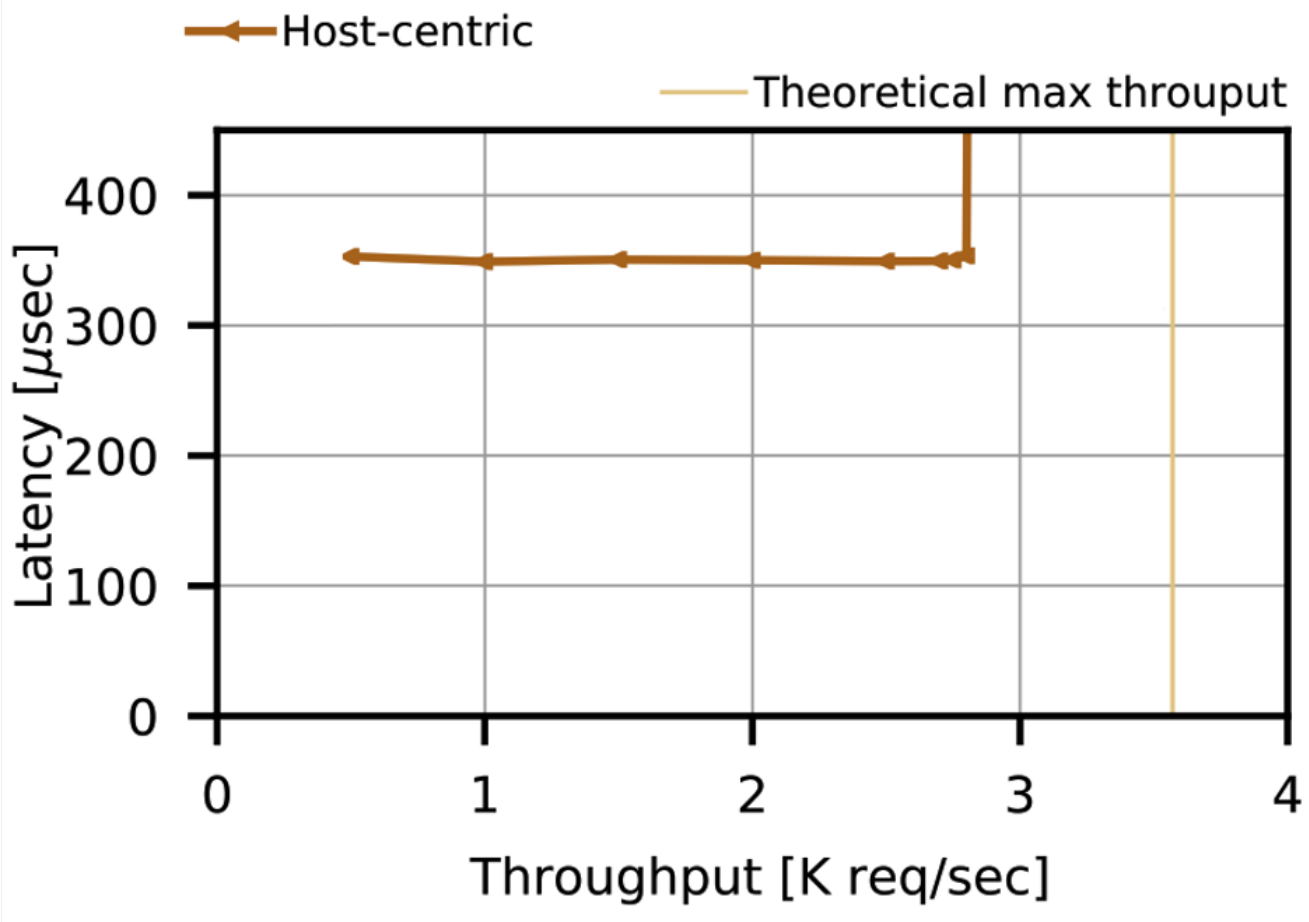
LeNet – Convolutional Neural Network Inference Server

LeNet is a DNN model for recognizing hand-written digits

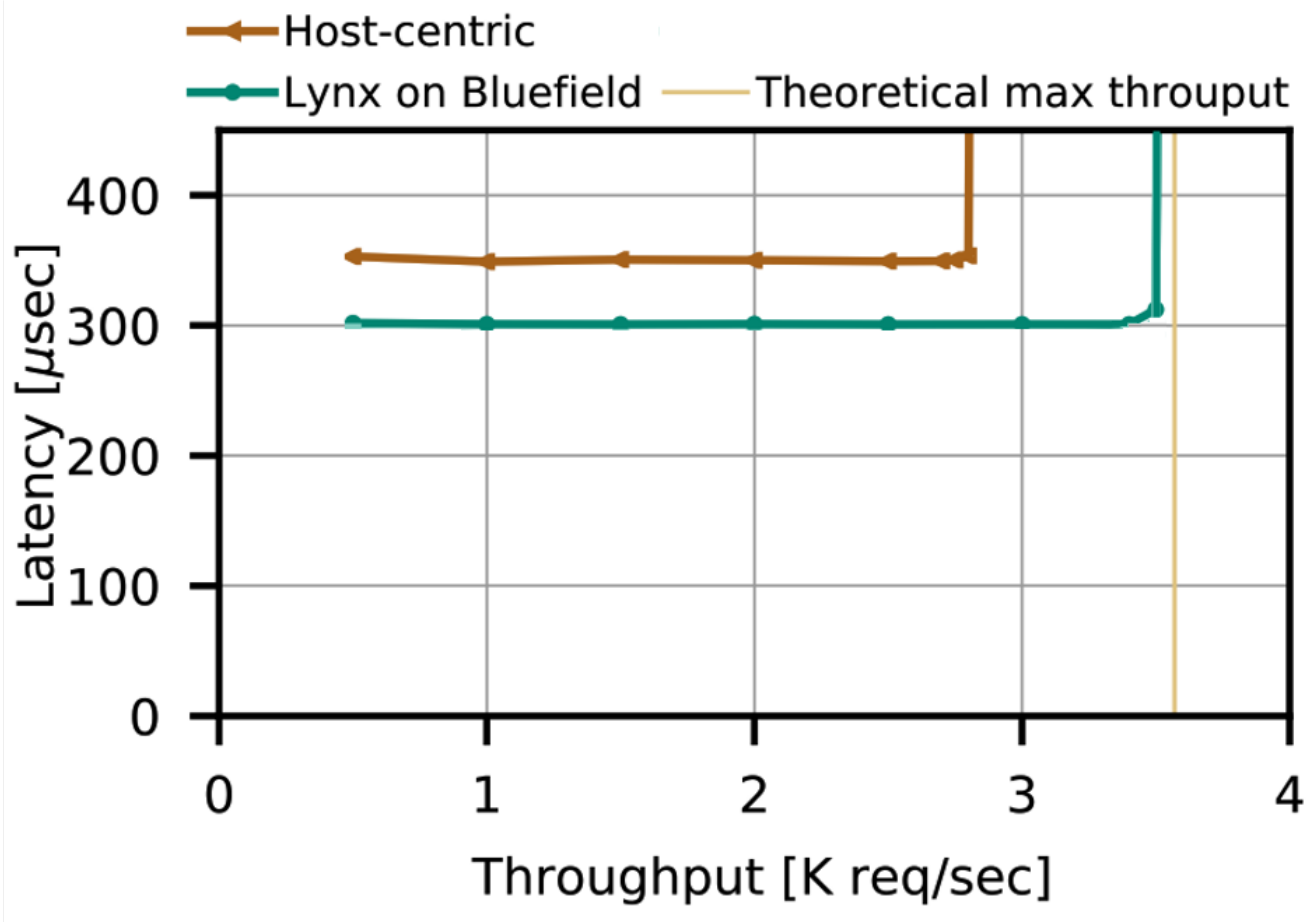
- Developed using **TensorFlow**
- Optimized using the **TVM compiler** with few modifications.
- No CPU code
- No application-specific SmartNIC code.



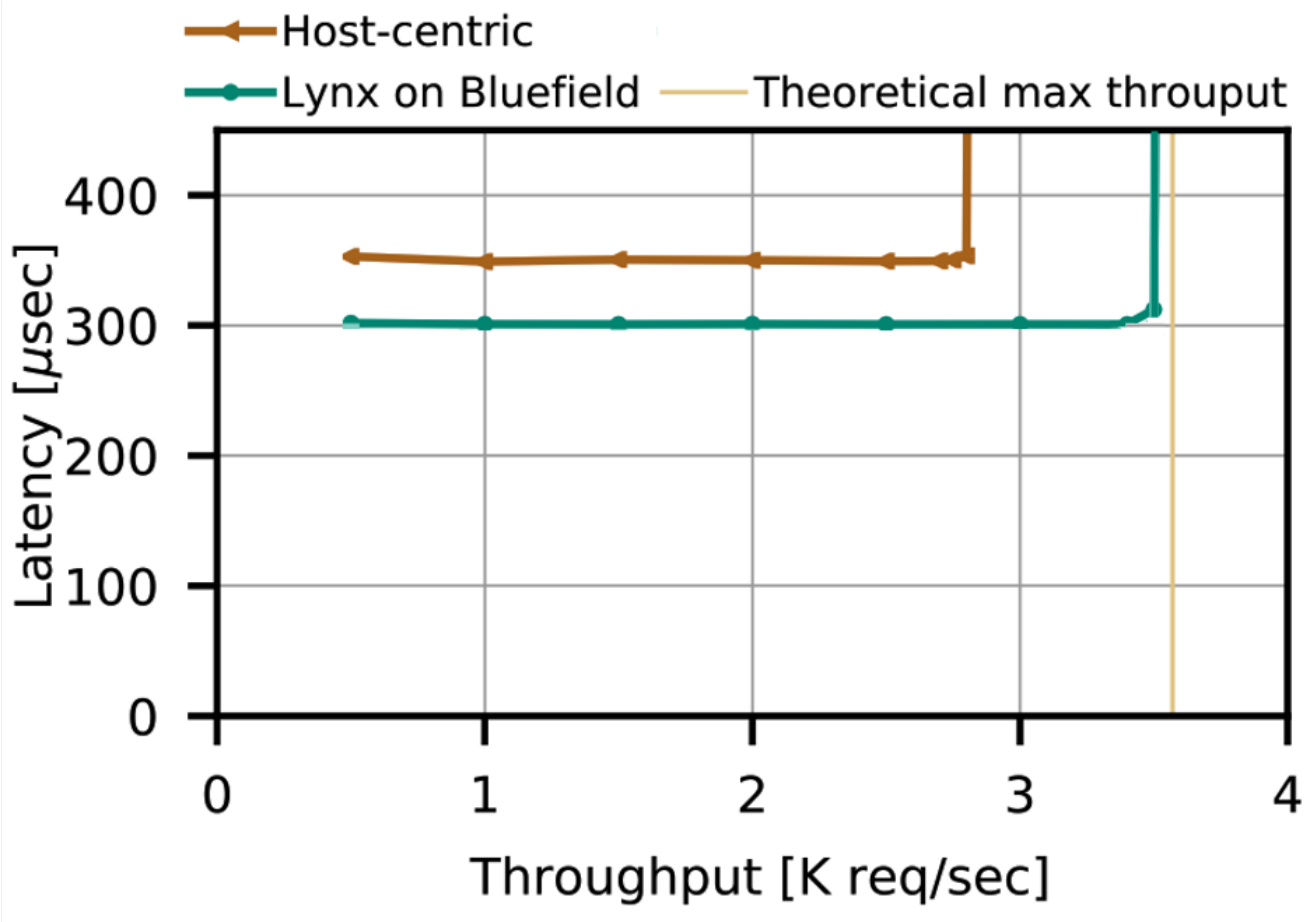
Host-centric is inefficient



Lynx achieves max theoretical throughput

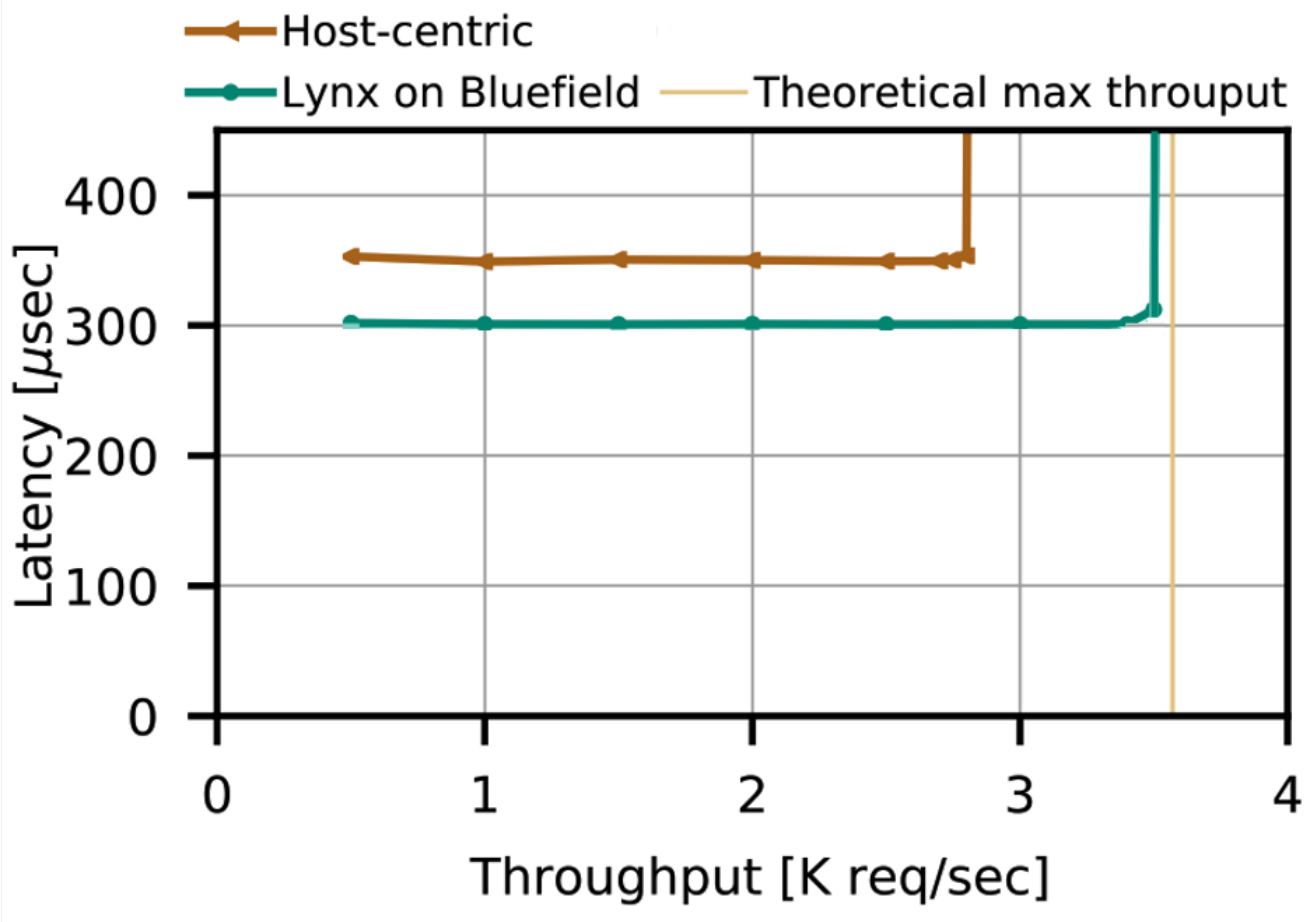


Lynx achieves max theoretical throughput



Throughput 25% ↑
Latency 20% ↓

Lynx achieves max theoretical throughput



Throughput 25% ↑

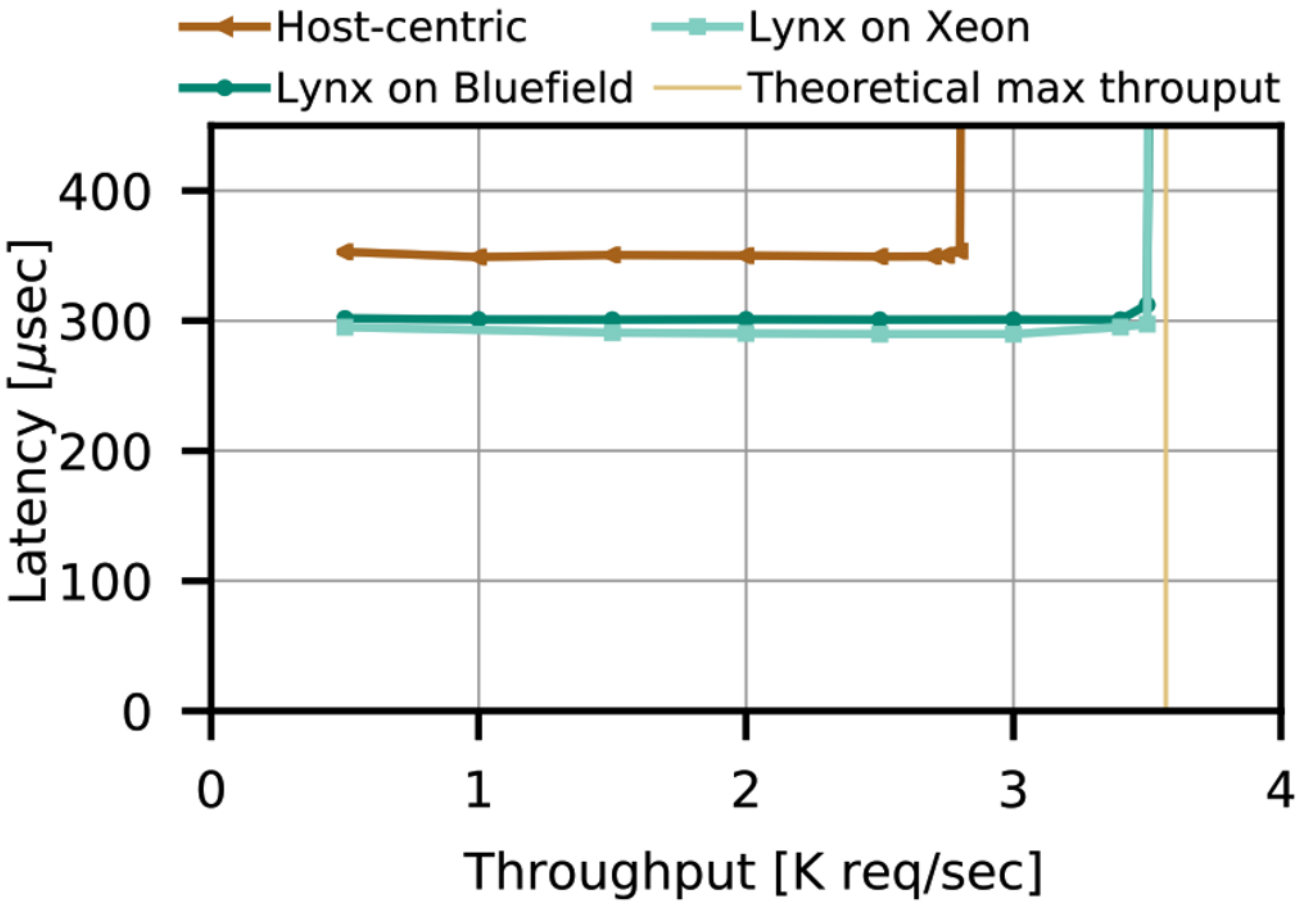
Latency 20% ↓

0 CPU Utilization

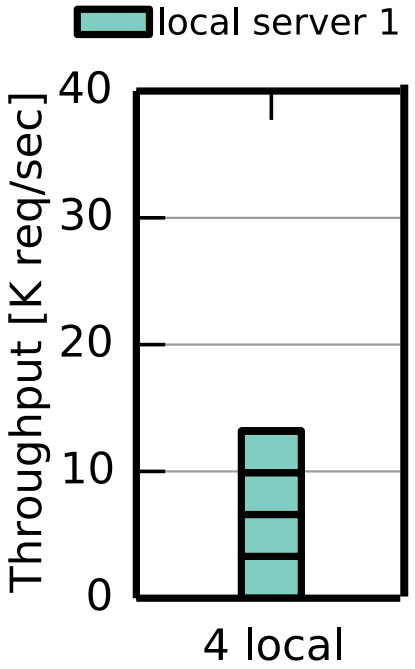
Lynx on BlueField and Xeon achieves the same performance

Takeaway

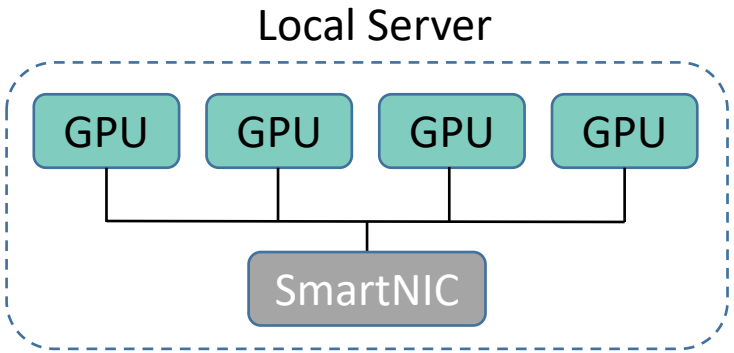
Comparing to Xeon core, Lynx on Bluefield can achieve the same throughput with negligible latency overhead.



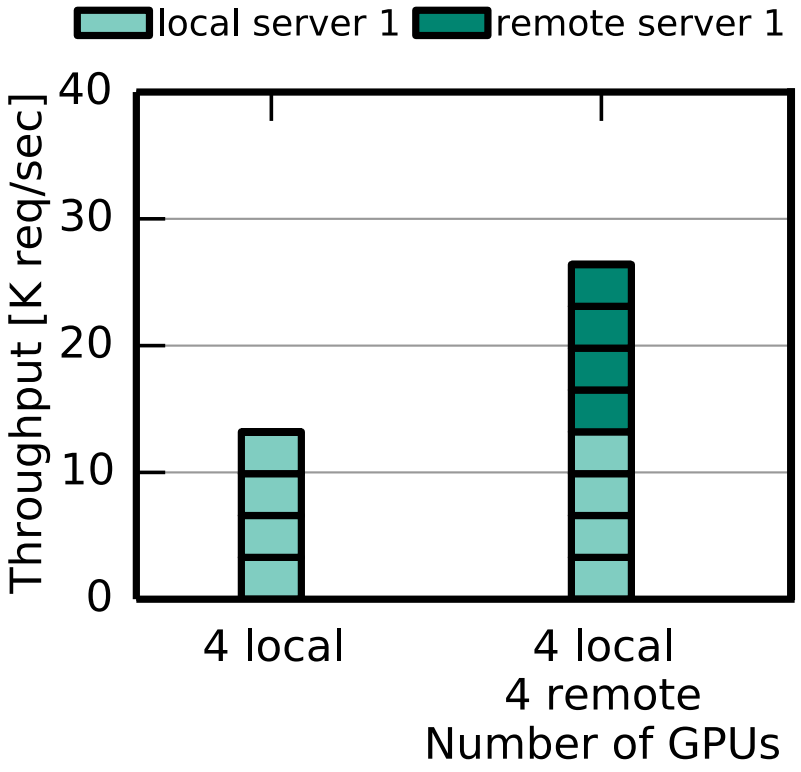
Inference server: Scalability with local GPUs

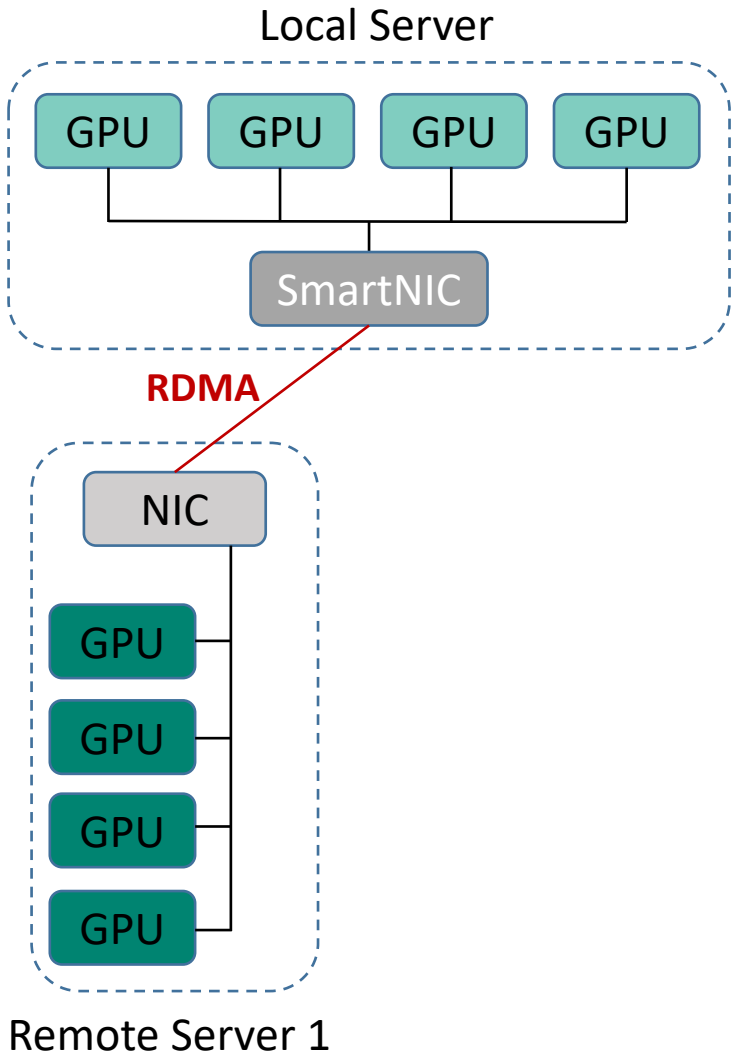


Number of GPUs

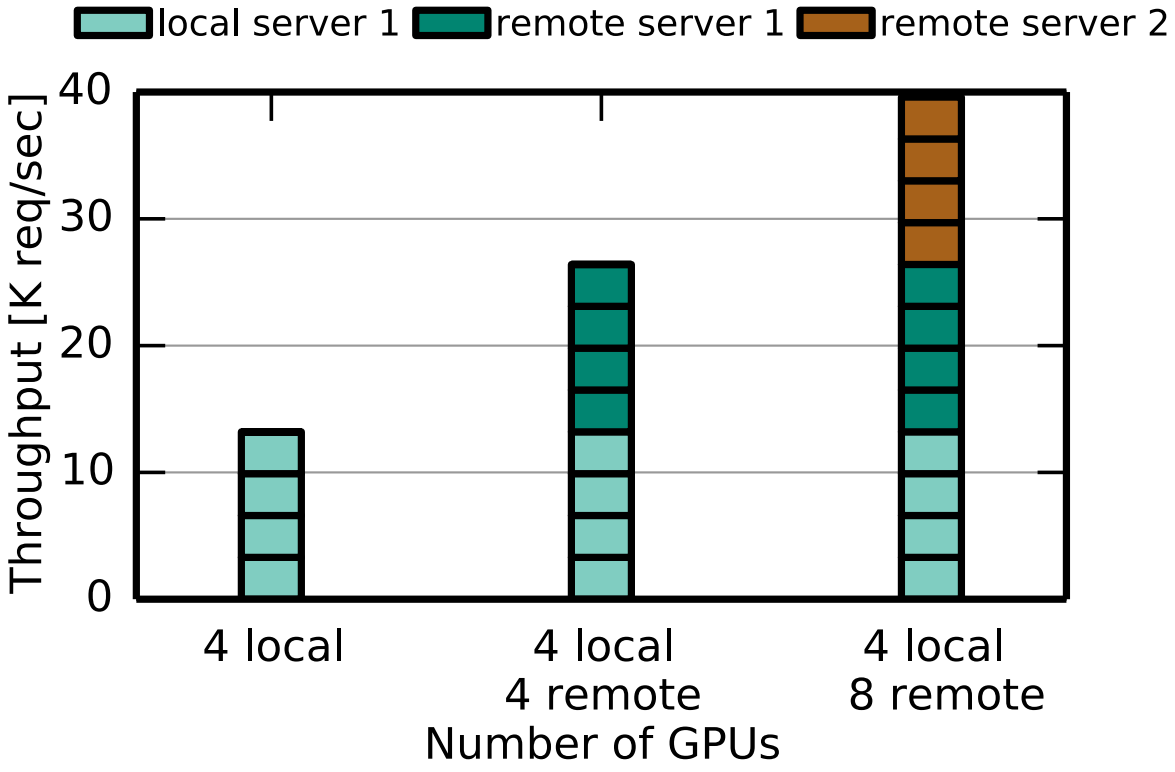


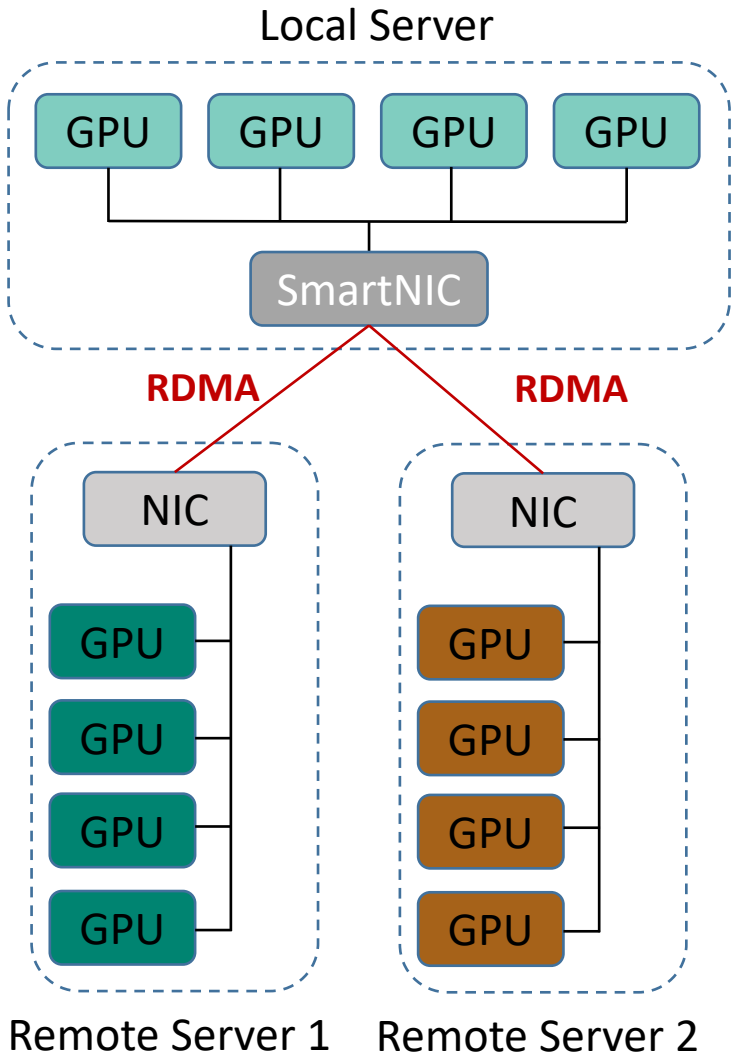
Inference server: Scalability with disaggregated GPUs



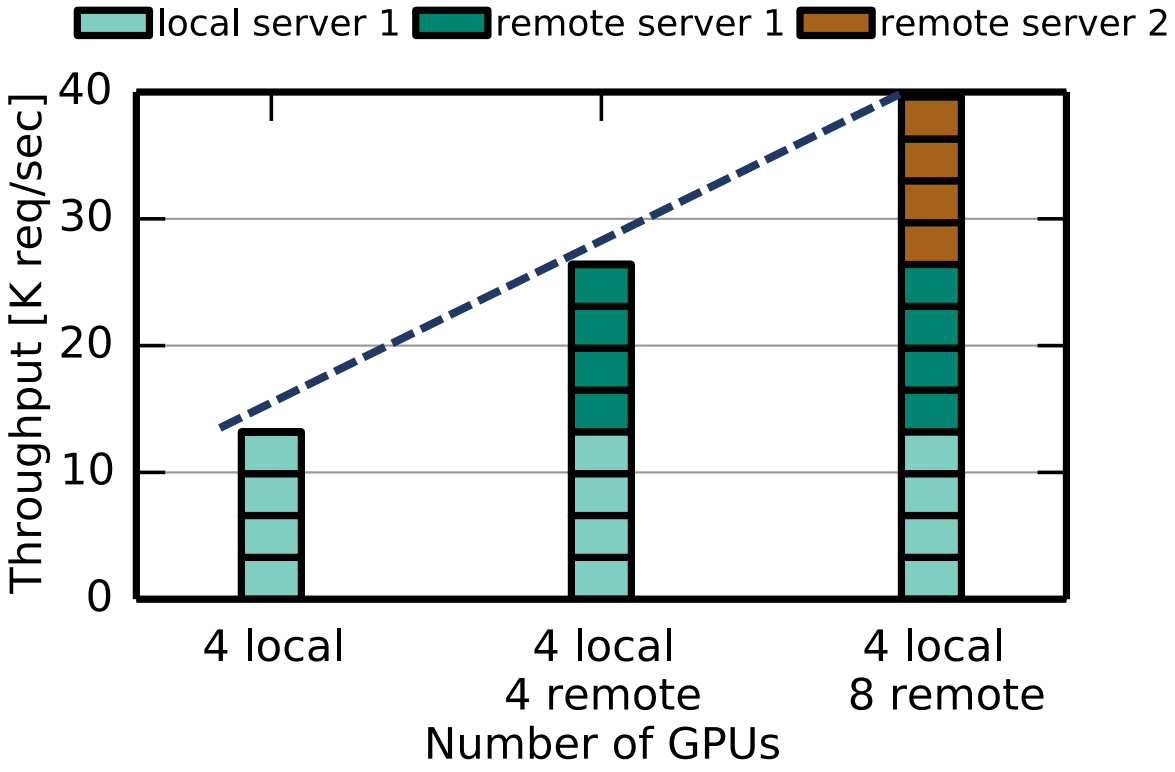


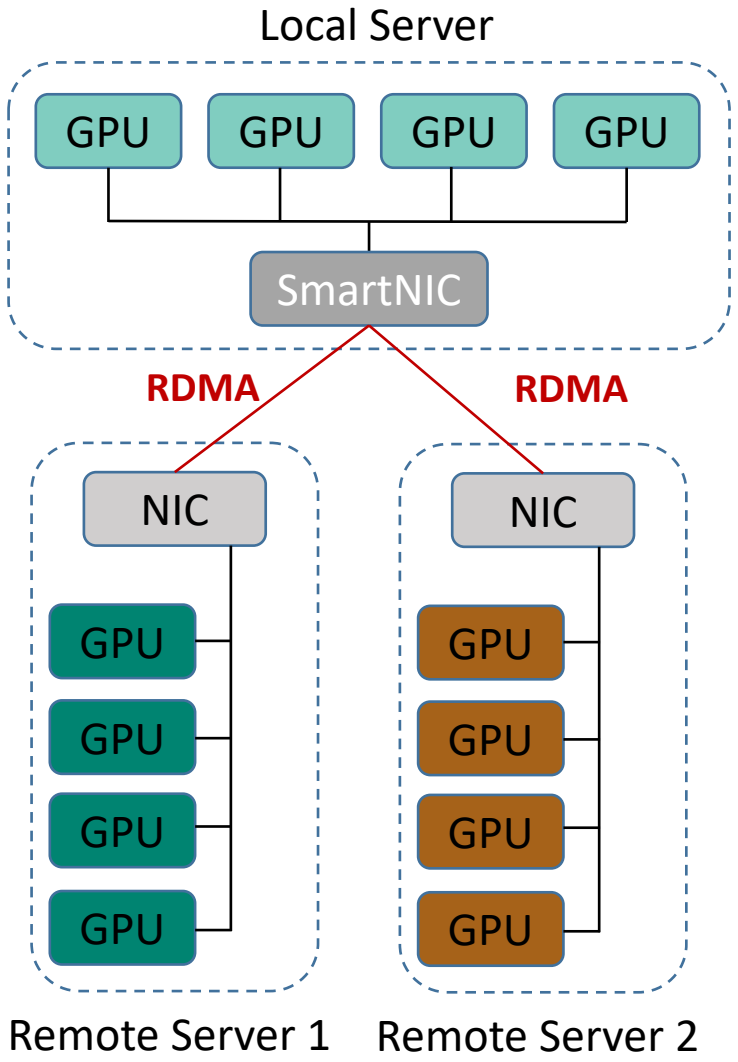
Inference server: Scalability with disaggregated GPUs





Inference server: Scalability with disaggregated GPUs

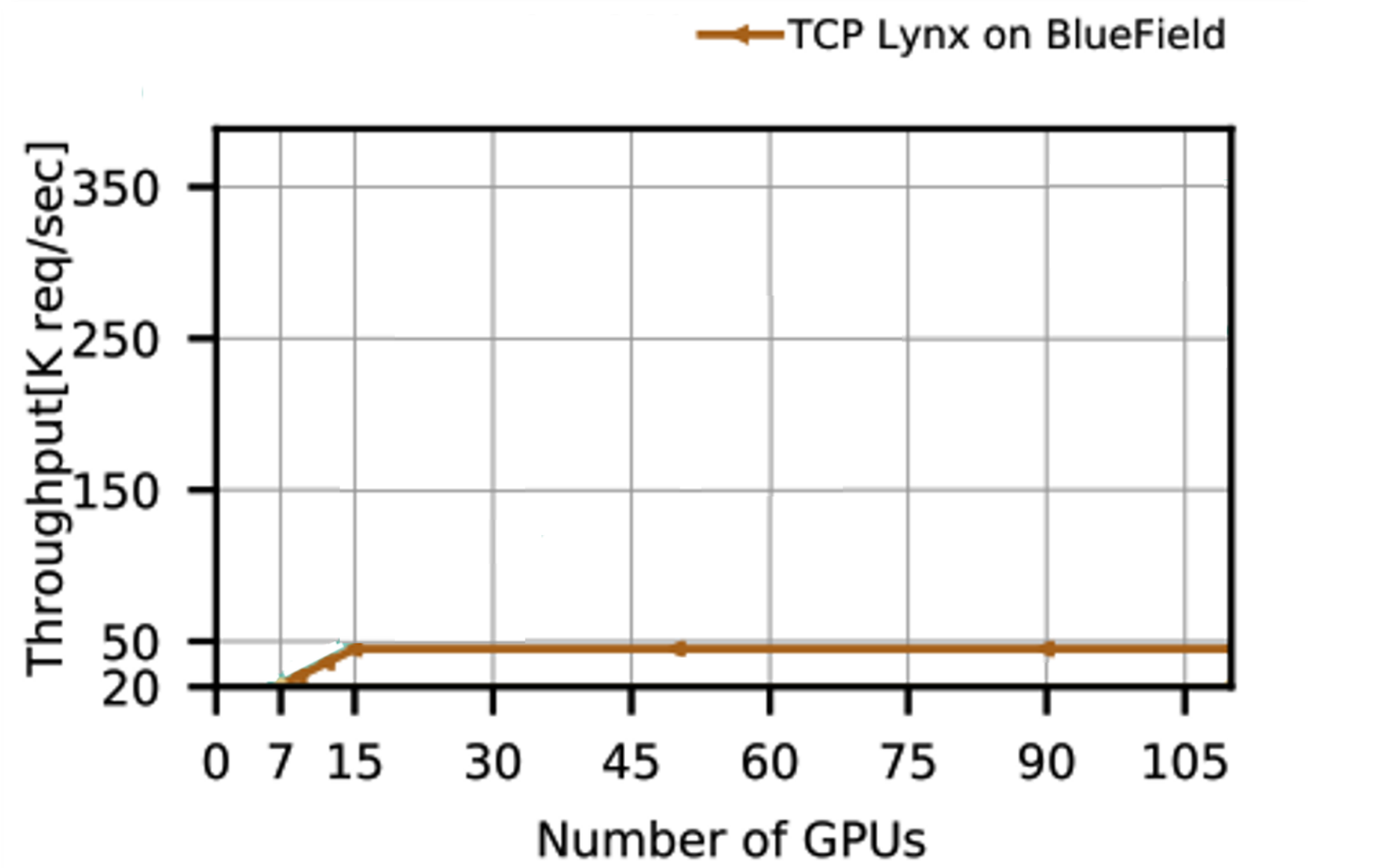




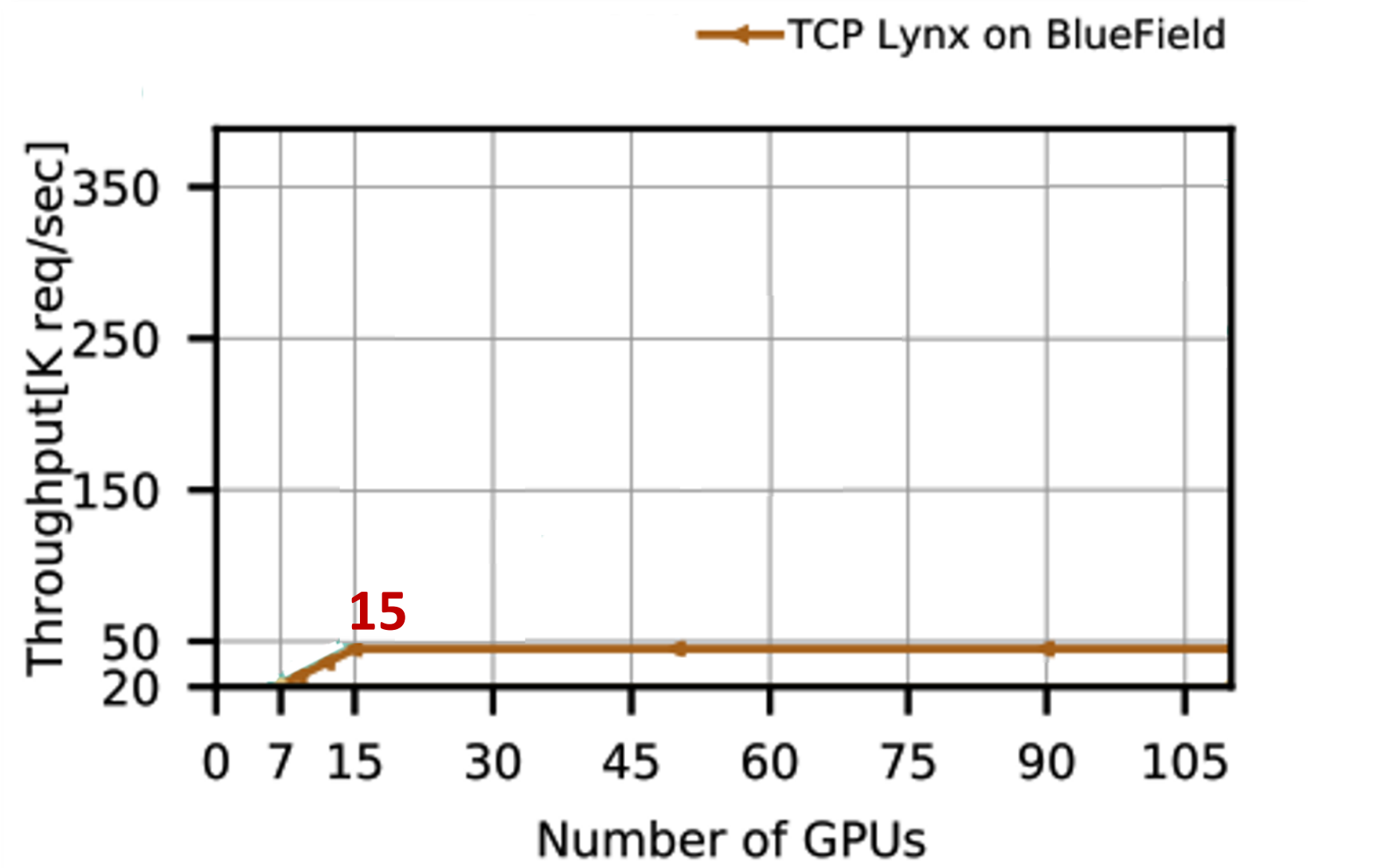
Takeaway

For compute bound applications LeNet scales linearly.

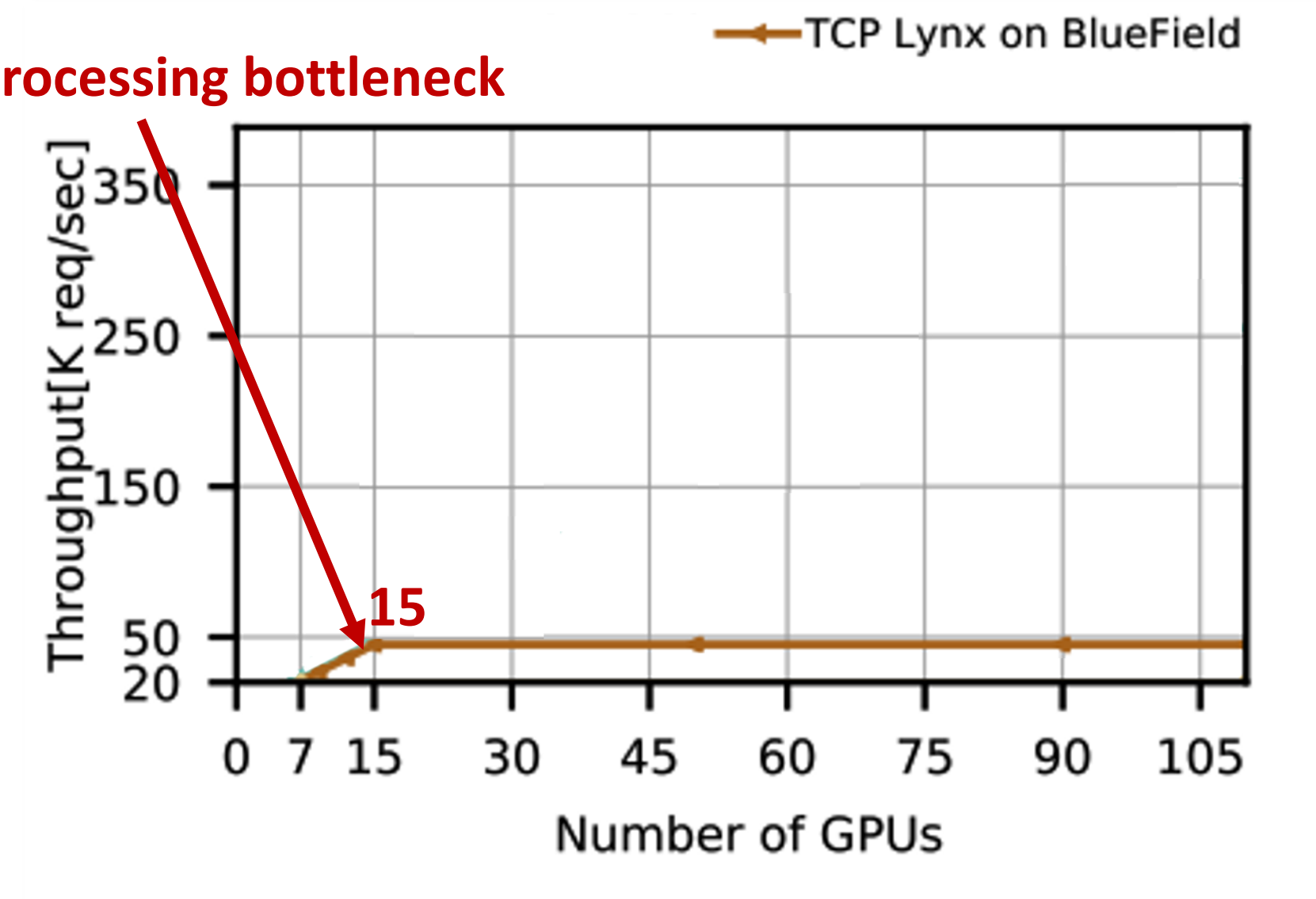
Scalability Projection (Upper bound)



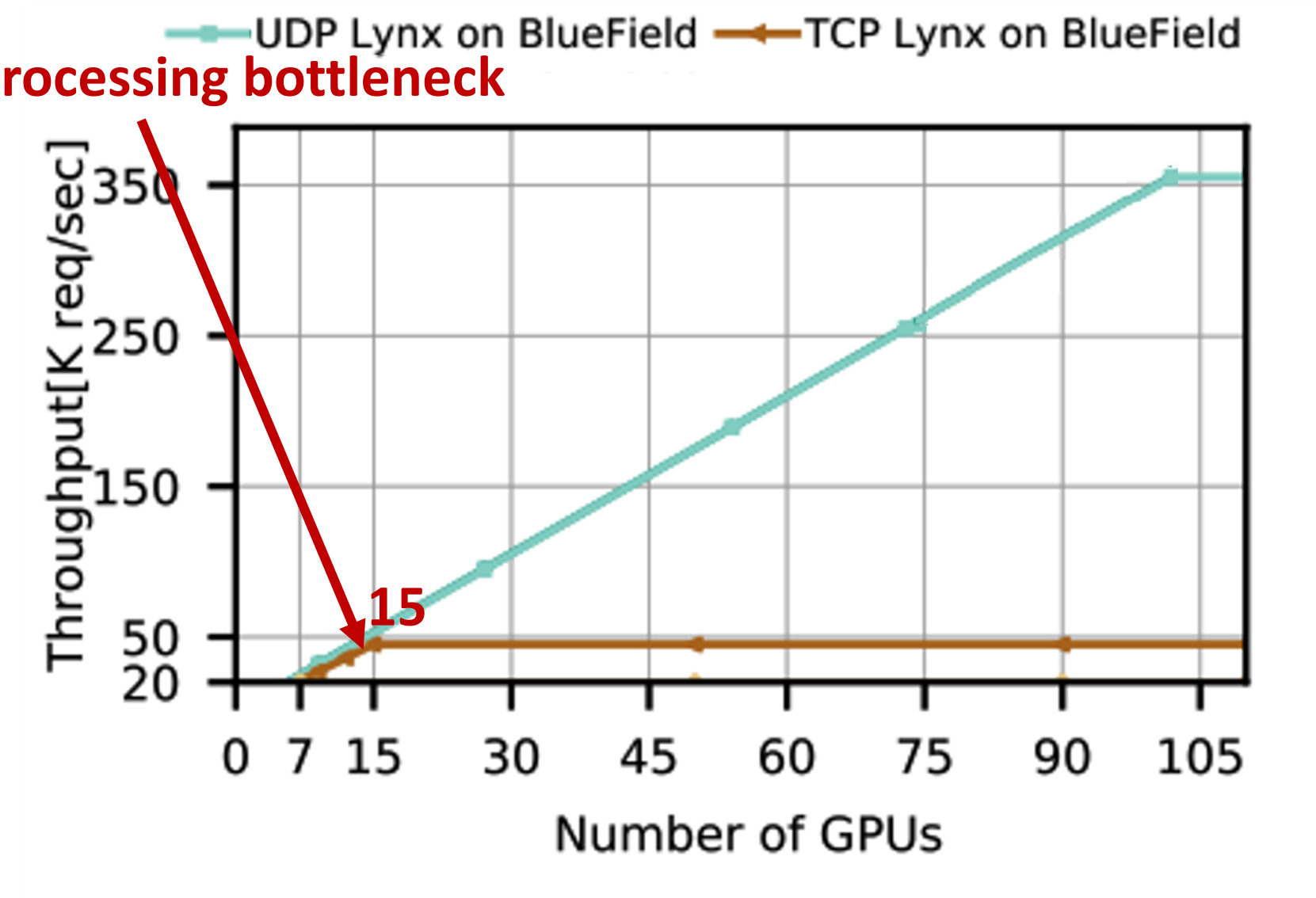
Scalability Projection (Upper bound)



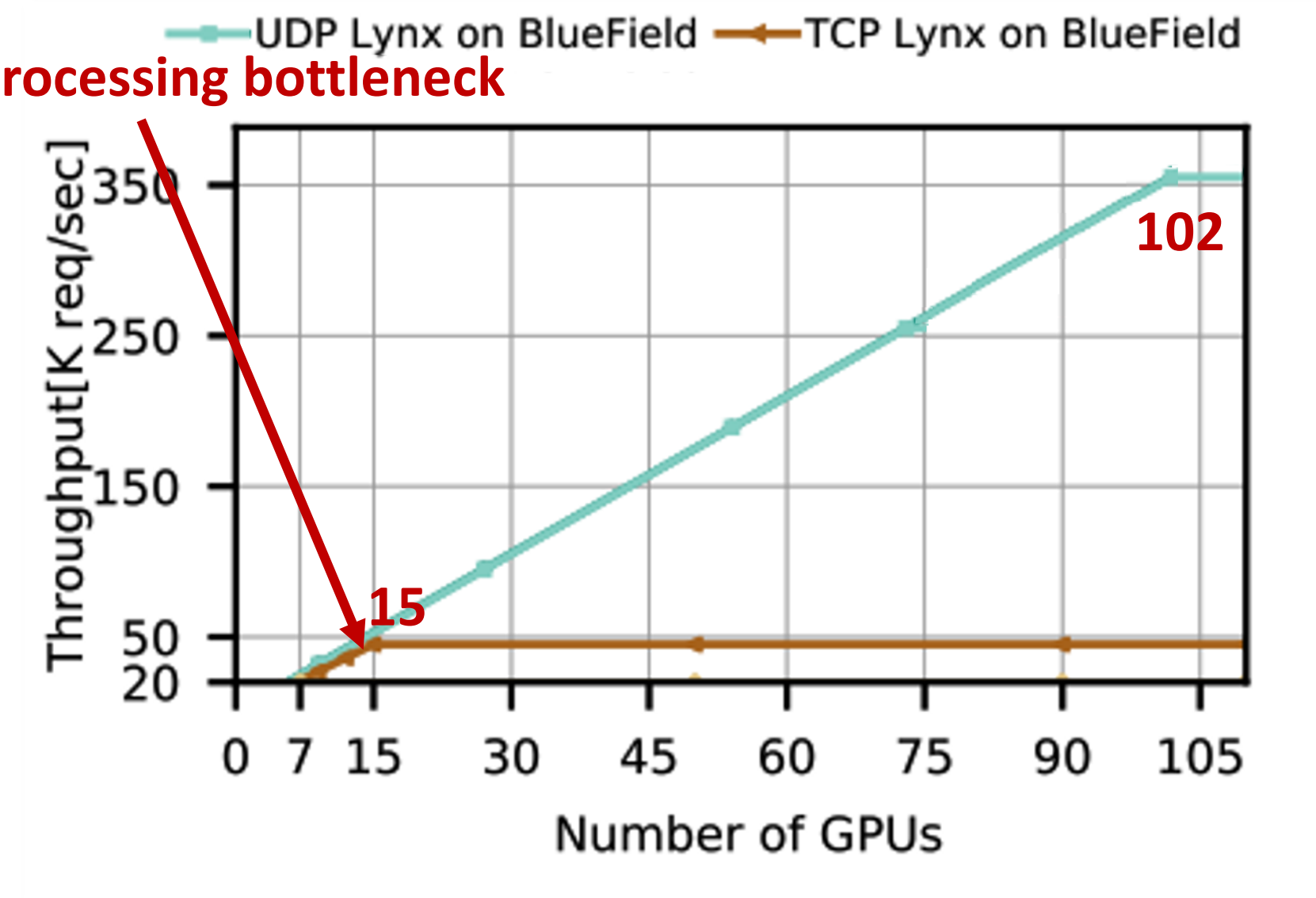
Scalability Projection (Upper bound)



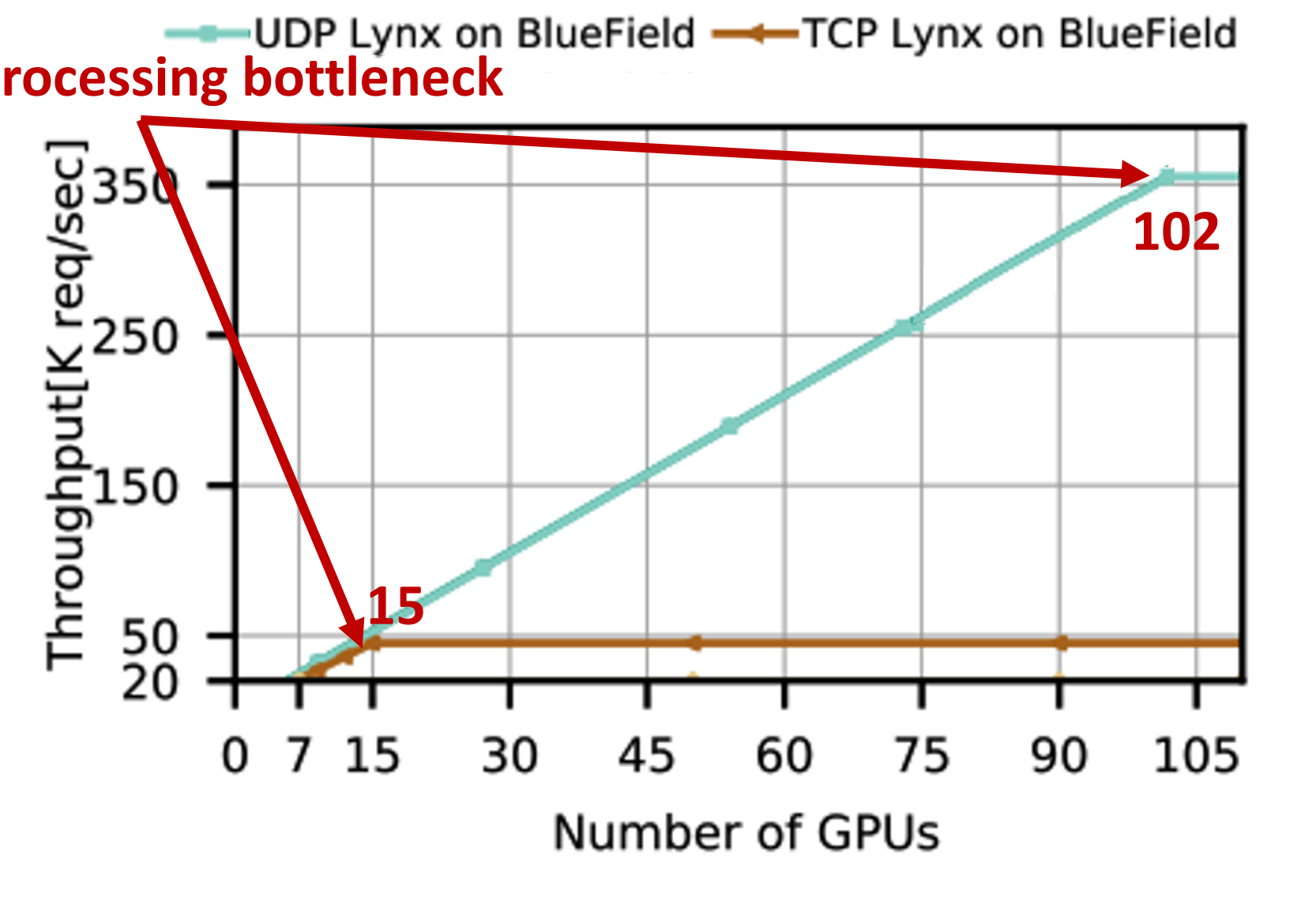
LeNet – Scalability Projection (Upper bound)



LeNet – Scalability Projection (Upper bound)



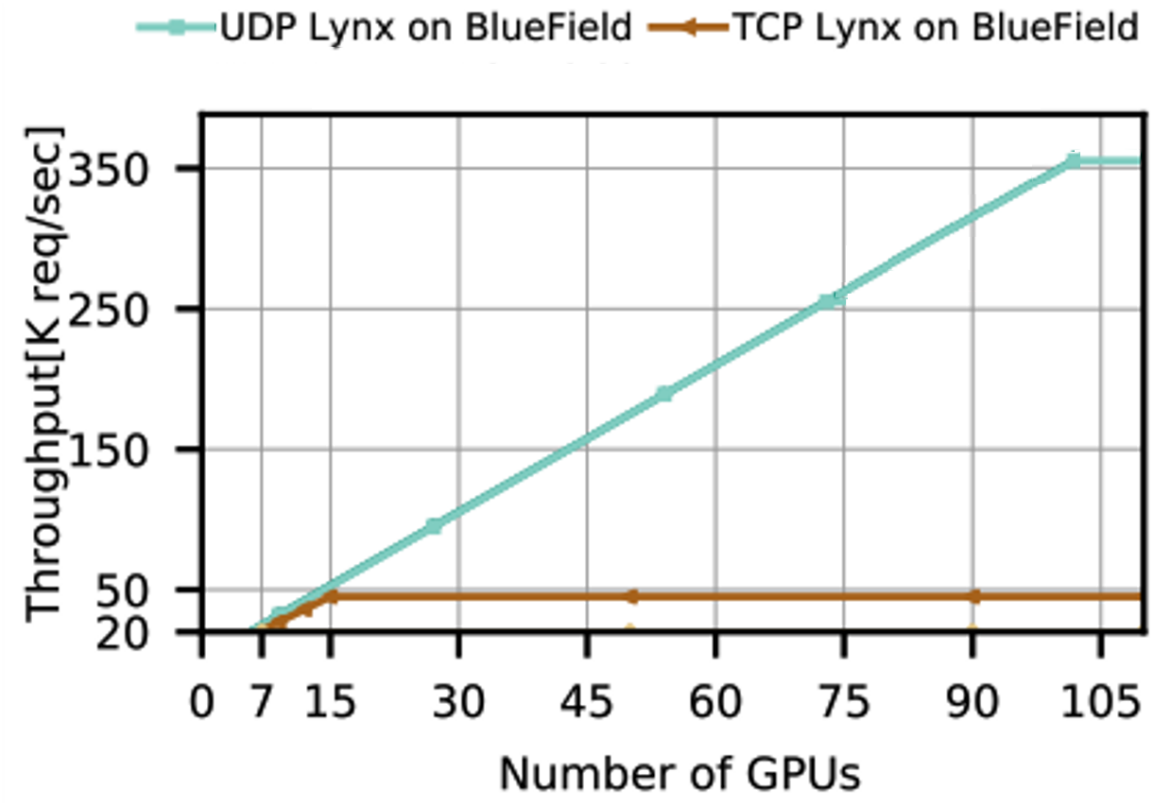
LeNet – Scalability Projection (Upper bound)



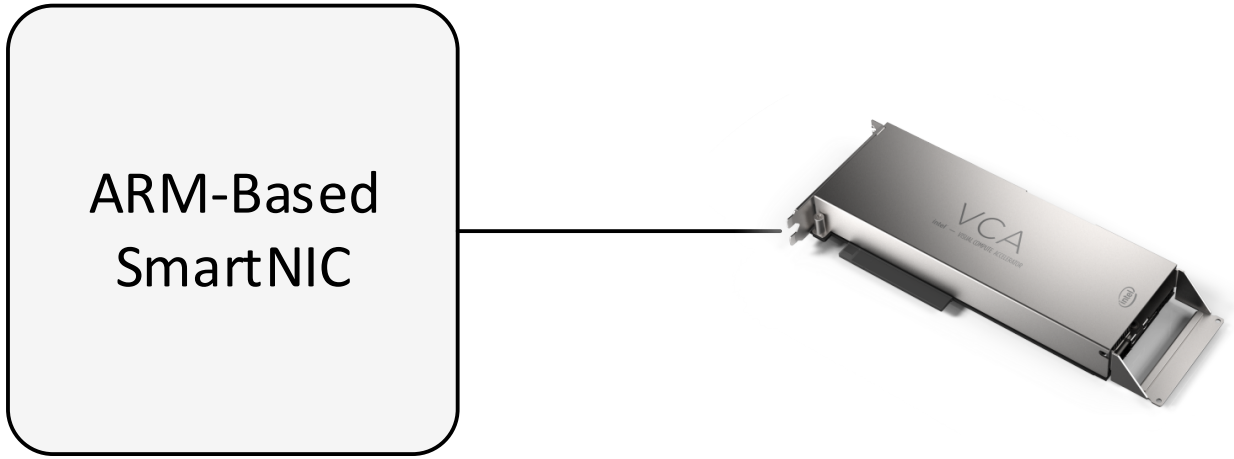
LeNet – Scalability Projection (Upper bound)

Takeaway

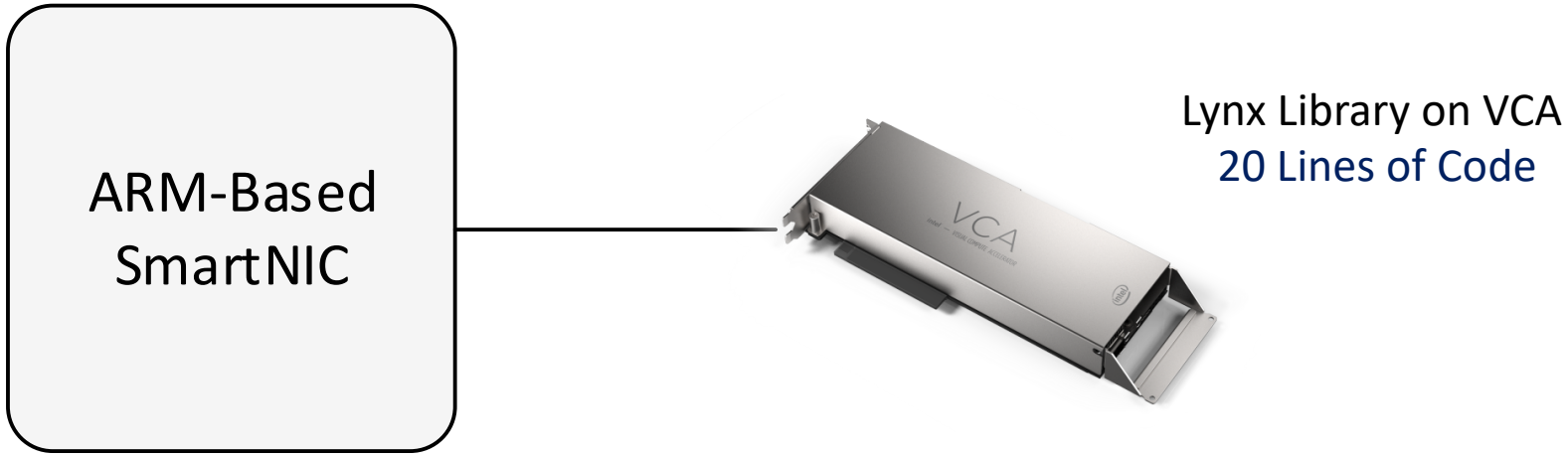
Lynx scales linearly until we reach the network processing bottleneck.



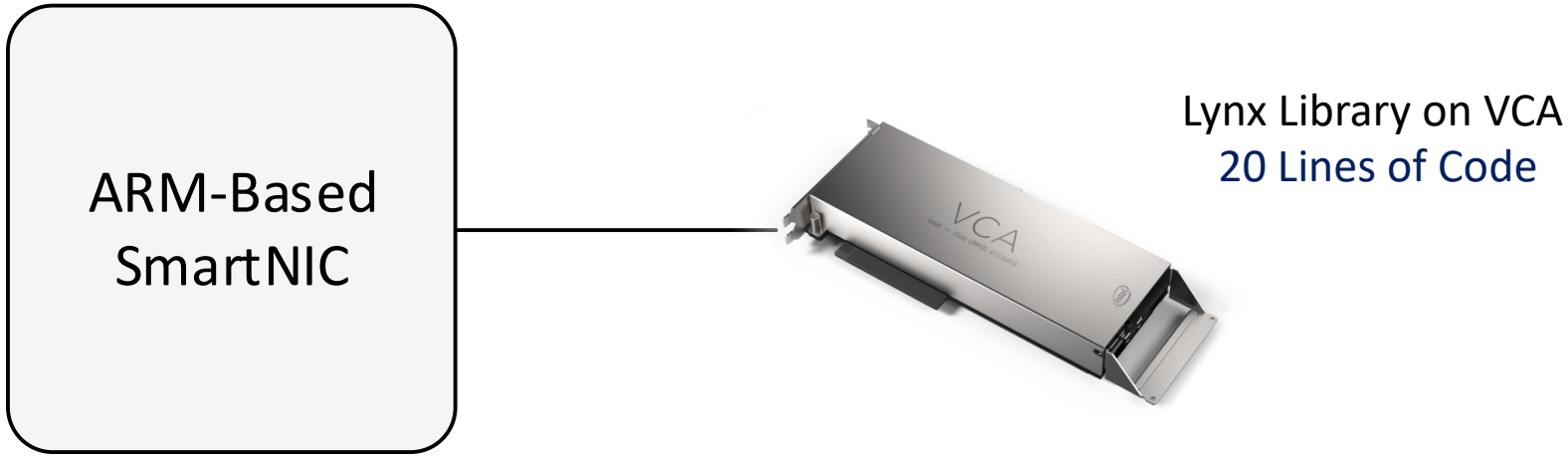
Portability: Lynx with the Intel VCA
Secure Computing Server inside SGX enclave



Portability: Lynx with the Intel VCA
Secure Computing Server inside SGX enclave



Portability: Lynx with the Intel VCA
Secure Computing Server inside SGX enclave

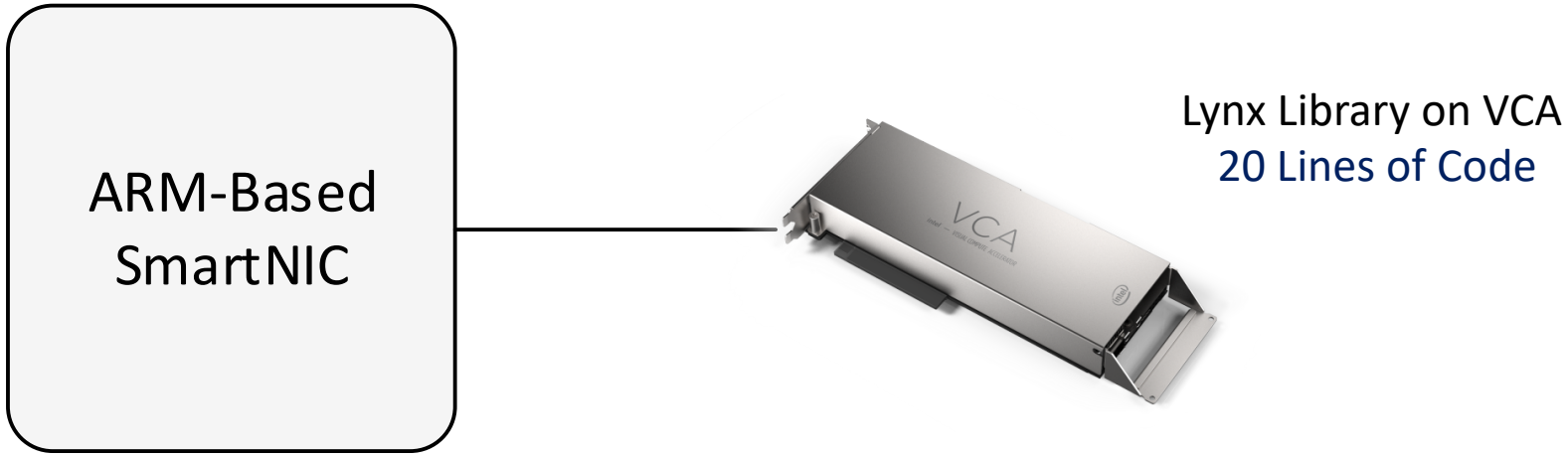


Lynx Library on VCA
20 Lines of Code

For each request, the VCA:

1. Decrypts an AES-encrypted message
2. Multiplies by a constant
3. Encrypts

Portability: Lynx with the Intel VCA
Secure Computing Server inside SGX enclave

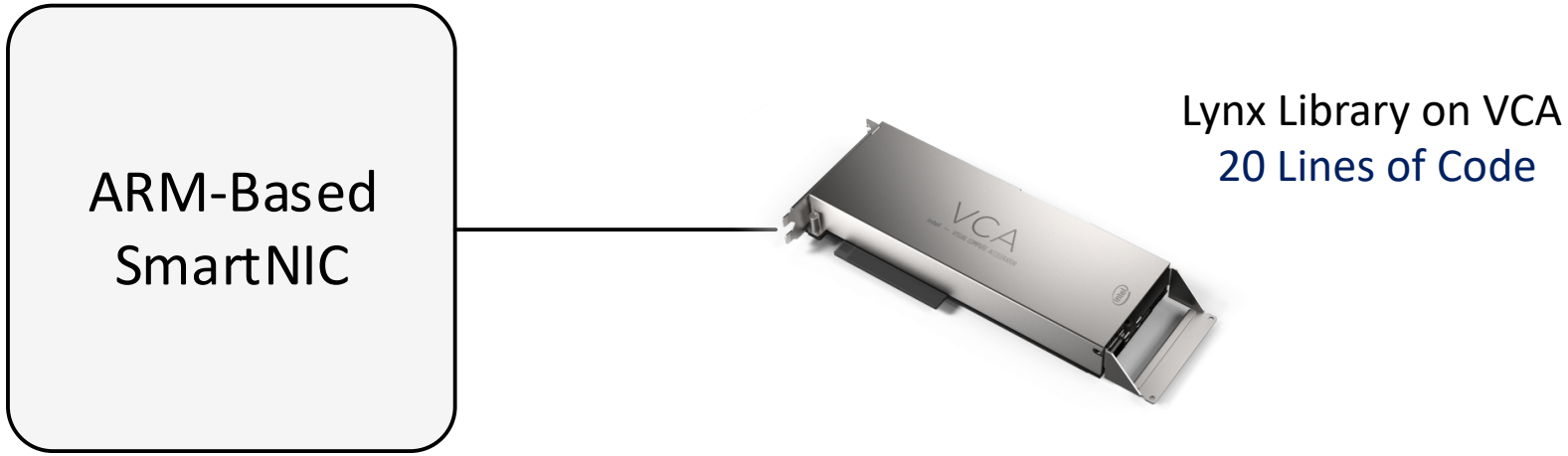


Lynx Library on VCA
20 Lines of Code

For each request, the VCA:
1. Decrypts an AES-encrypted message
2. Multiplies by a constant
3. Encrypts

SGX guarantees that the encryption key is not accessible from the server.

Portability: Lynx with the Intel VCA
Secure Computing Server inside SGX enclave



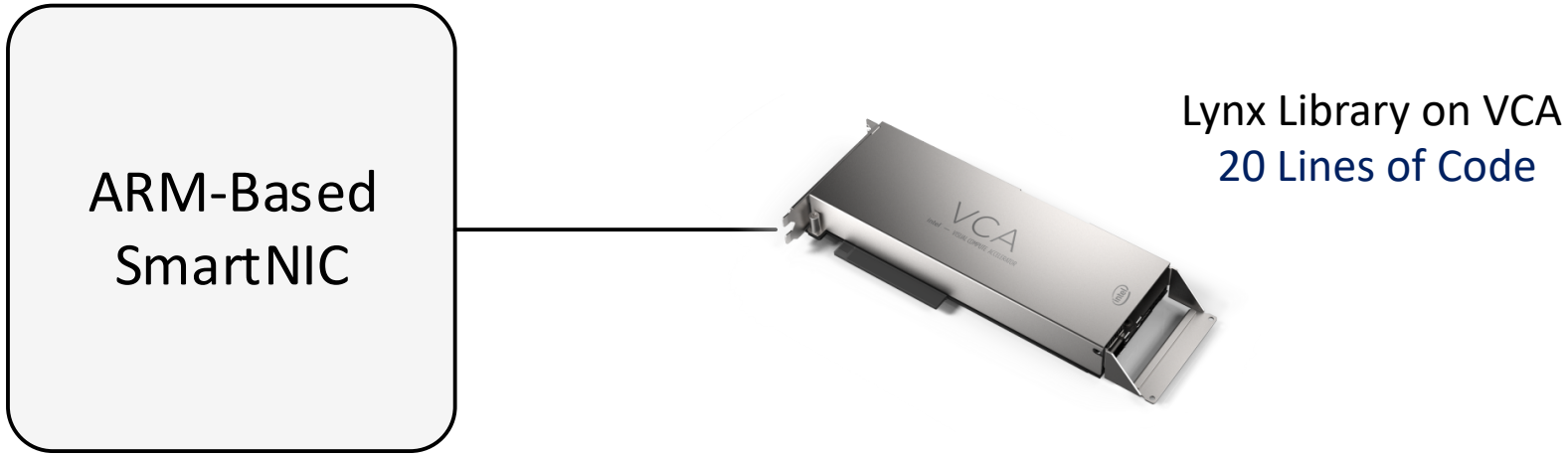
Lynx Library on VCA
20 Lines of Code

For each request, the VCA:
1. Decrypts an AES-encrypted message
2. Multiplies by a constant
3. Encrypts

SGX guarantees that the encryption key is not accessible from the server.

	99^{th} $percentile$
Host-Centric	259.14 μsec
Lynx	57.5 μsec

Portability: Lynx with the Intel VCA
Secure Computing Server inside SGX enclave



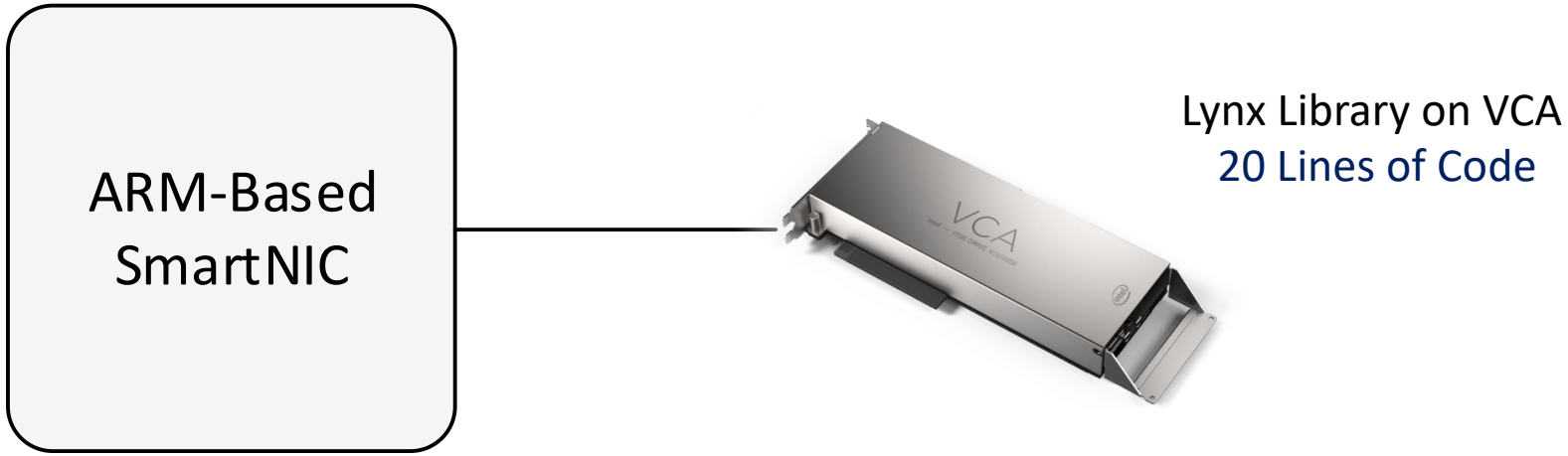
Lynx Library on VCA
20 Lines of Code

For each request, the VCA:
1. Decrypts an AES-encrypted message
2. Multiplies by a constant
3. Encrypts

SGX guarantees that the encryption key is not accessible from the server.

	99^{th} $percentile$
Host-Centric	259.14 μsec
Lynx	57.5 μsec

Portability: Lynx with the Intel VCA
Secure Computing Server inside SGX enclave



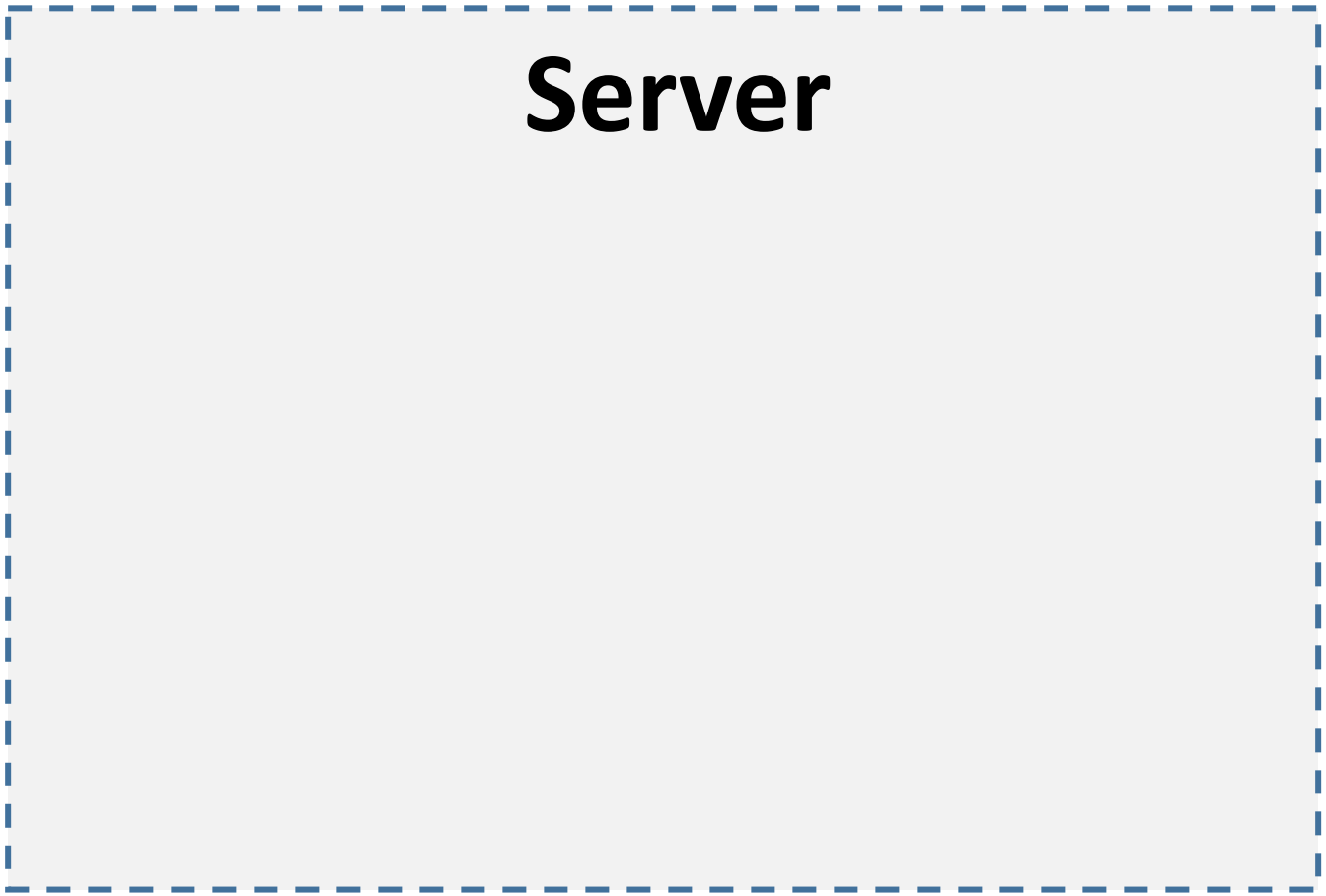
Takeaway

Integrating a new accelerator with Lynx is very simple.

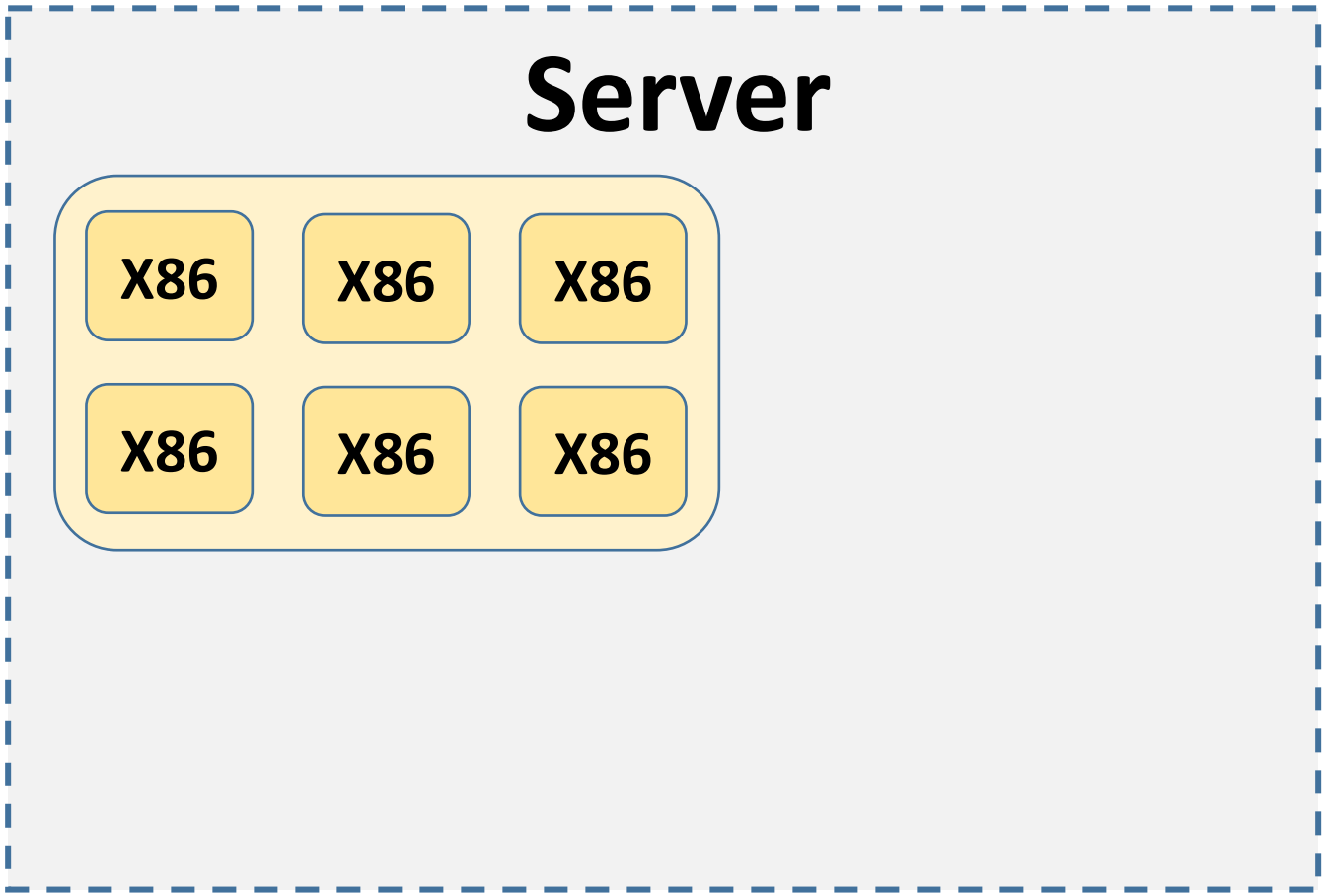
	99^{th} $percentile$
Host-Centric	259.14 μsec
Lynx	57.5 μsec

Why not using the SmartNIC as another CPU?

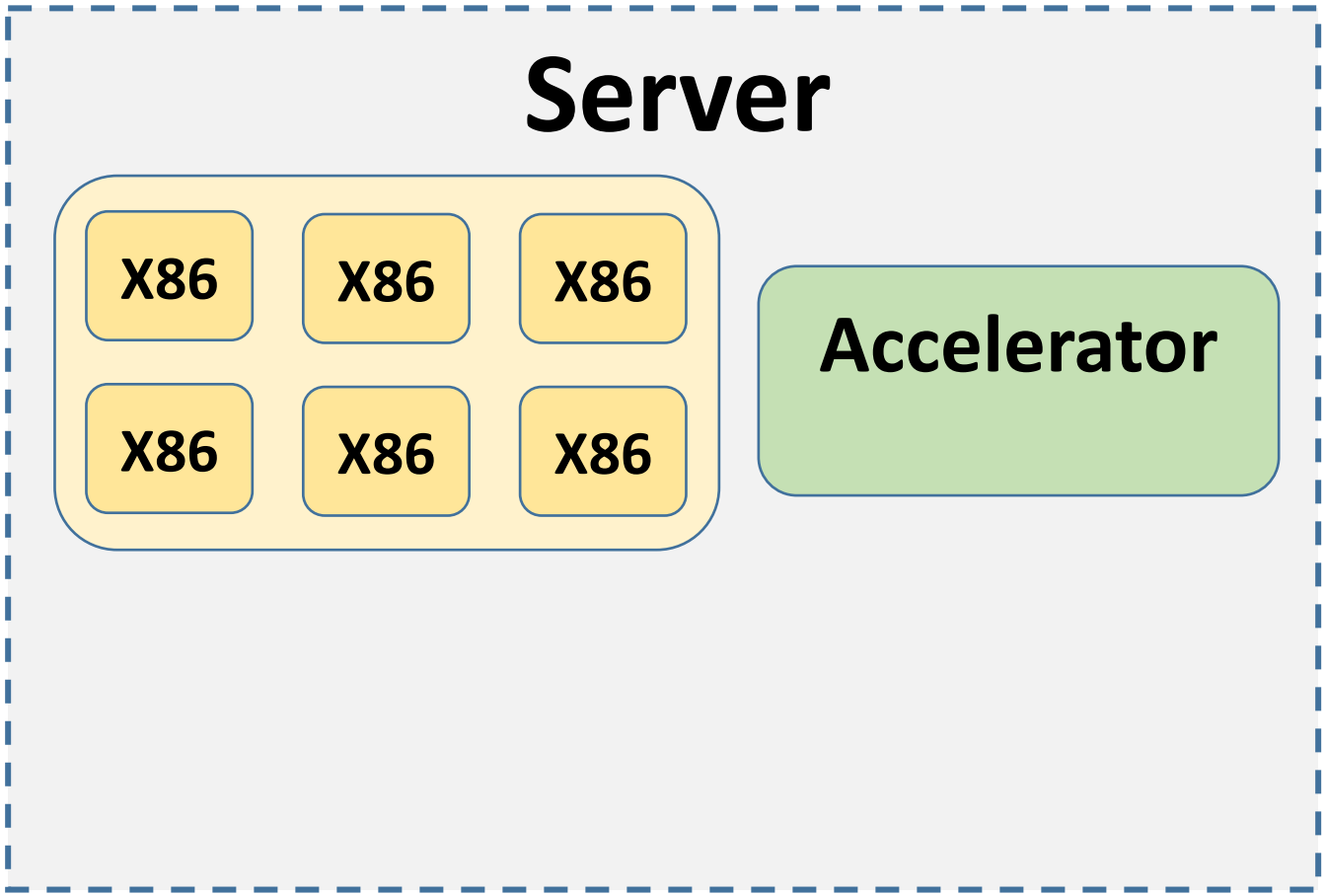
Why not using the SmartNIC as another CPU?



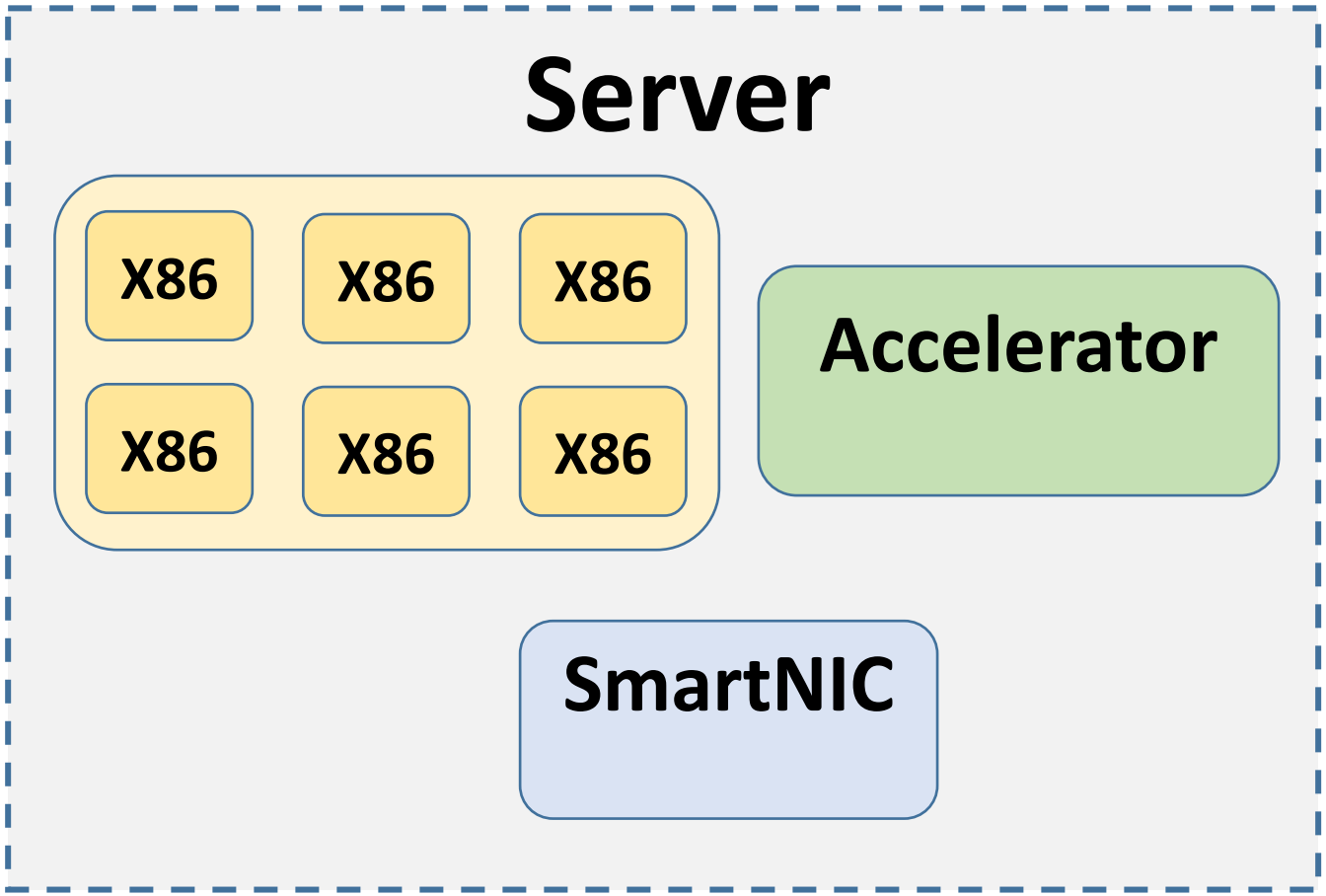
Why not using the SmartNIC as another CPU?



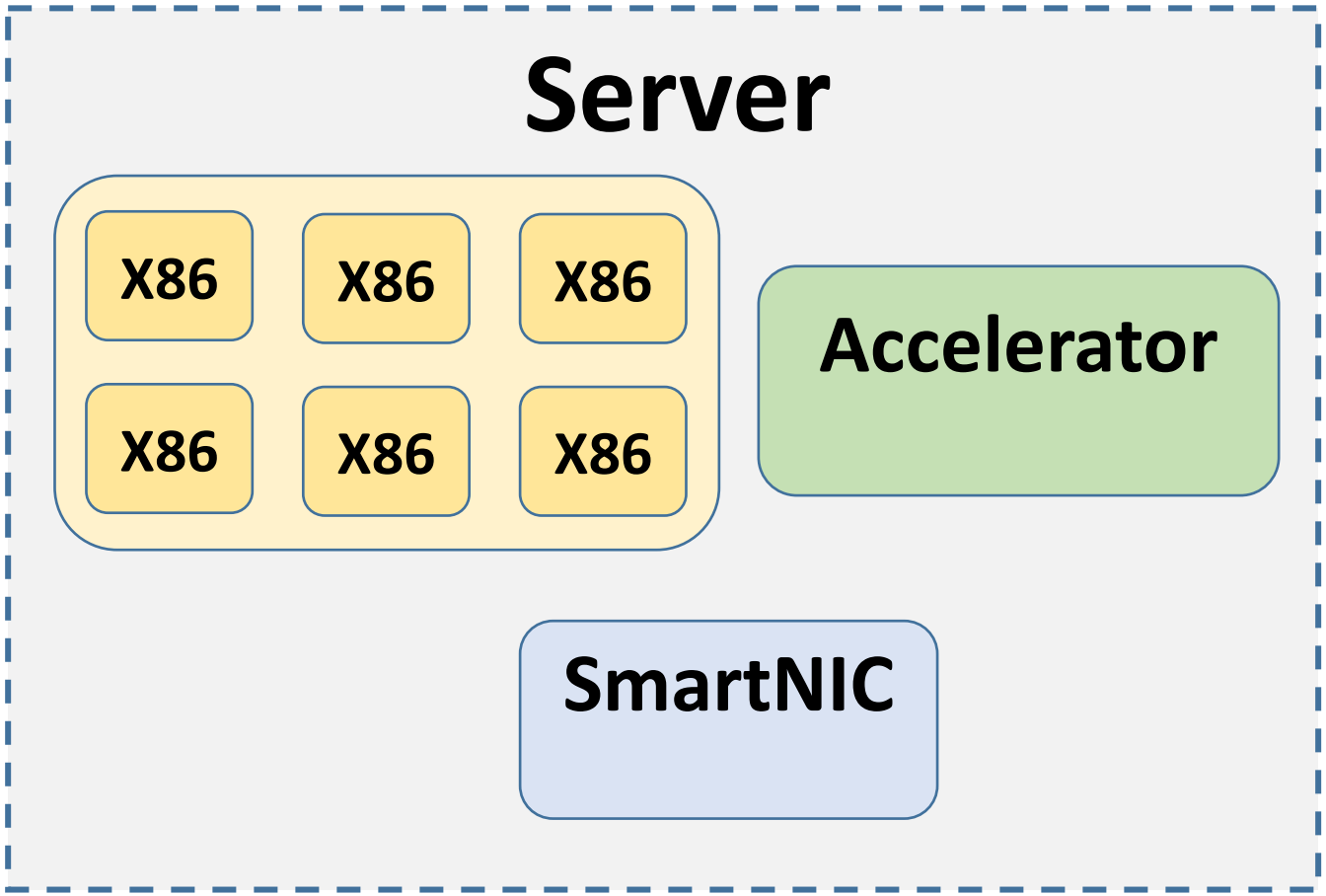
Why not using the SmartNIC as another CPU?



Why not using the SmartNIC as another CPU?



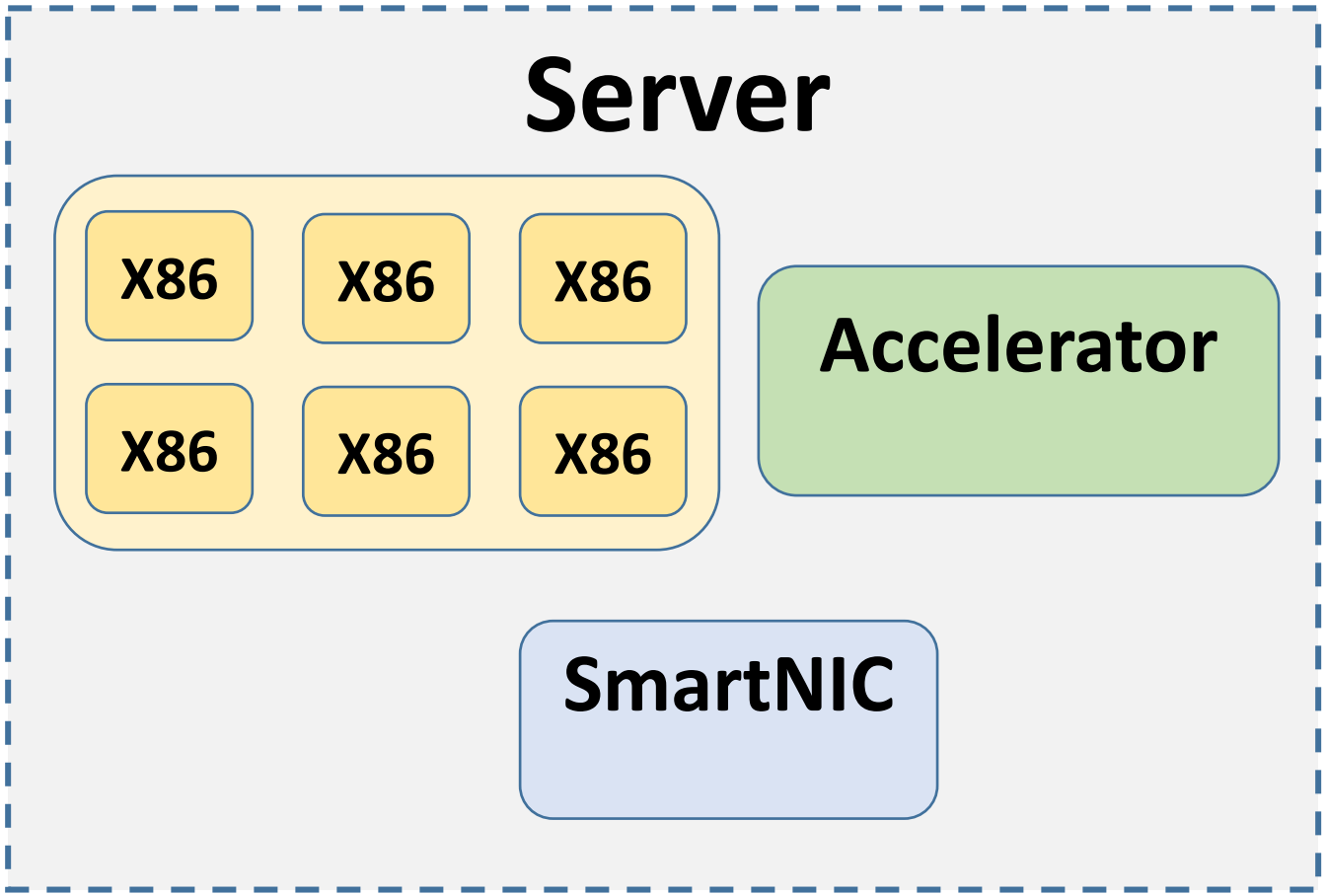
Why not using the SmartNIC as another CPU?



Applications

- memcached
- LeNet

Why not using the SmartNIC as another CPU?

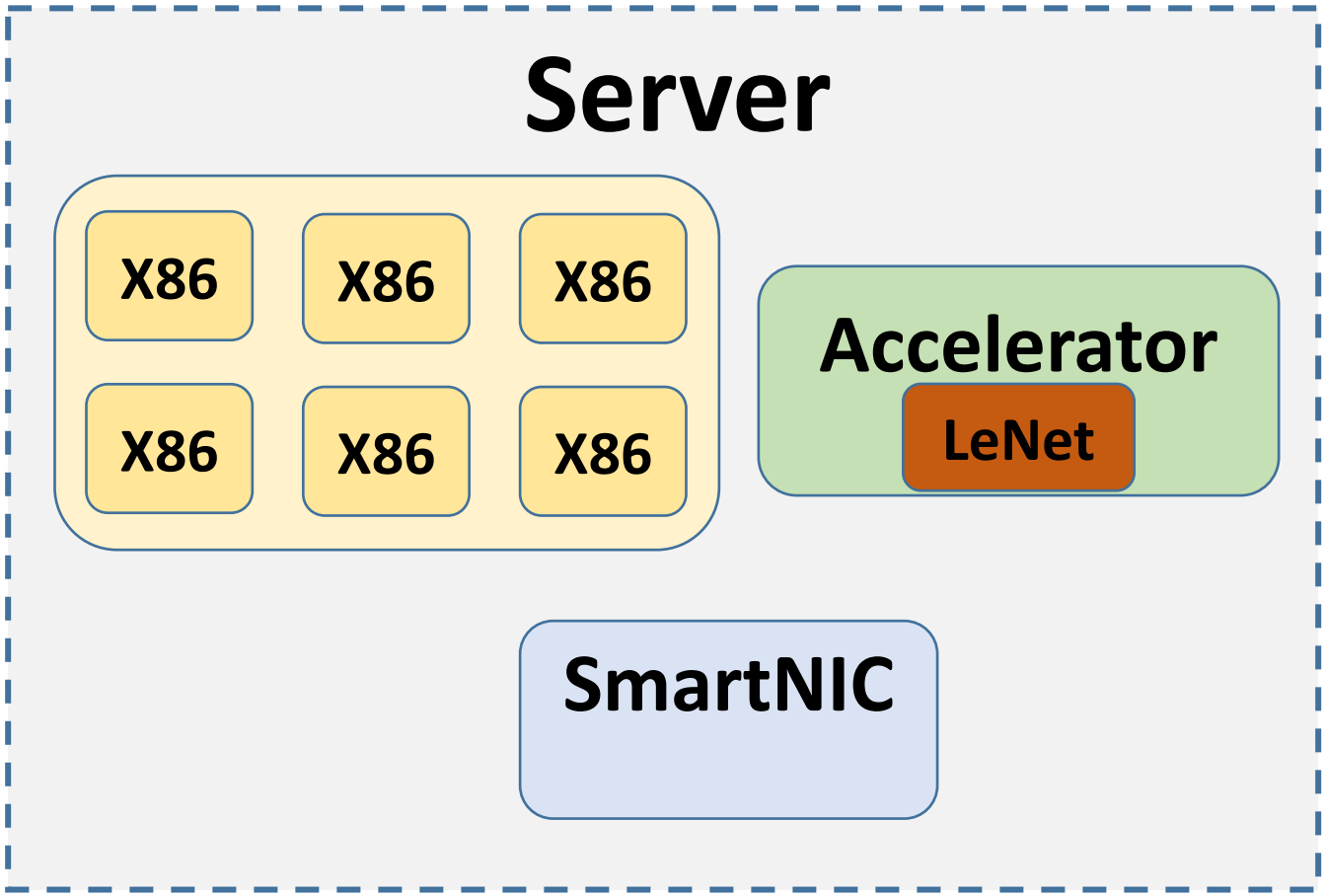


Applications

- memcached
- LeNet

What is the best mapping?

Why not using the SmartNIC as another CPU?

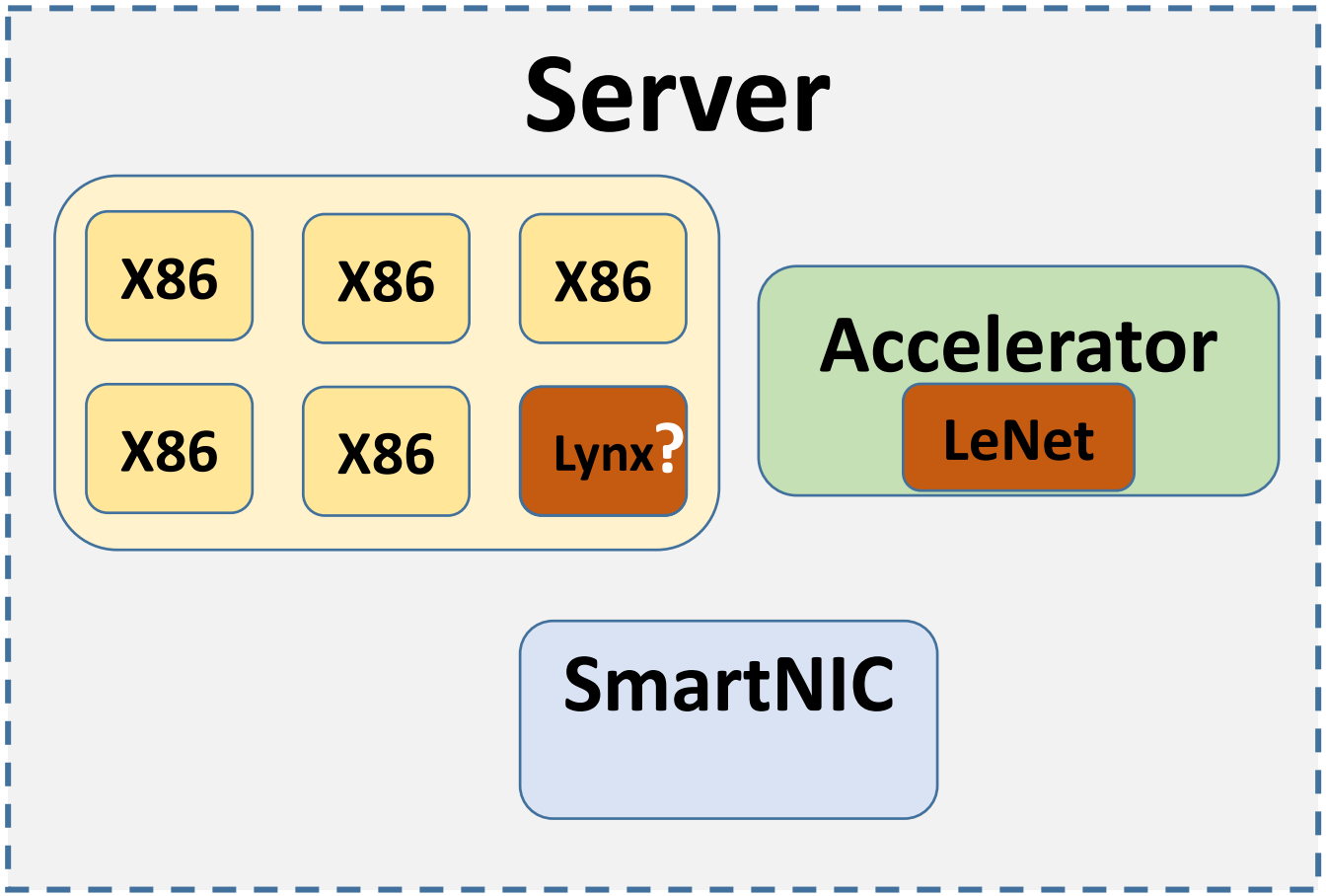


Applications

memcached

What is the best mapping?

Why not using the SmartNIC as another CPU?

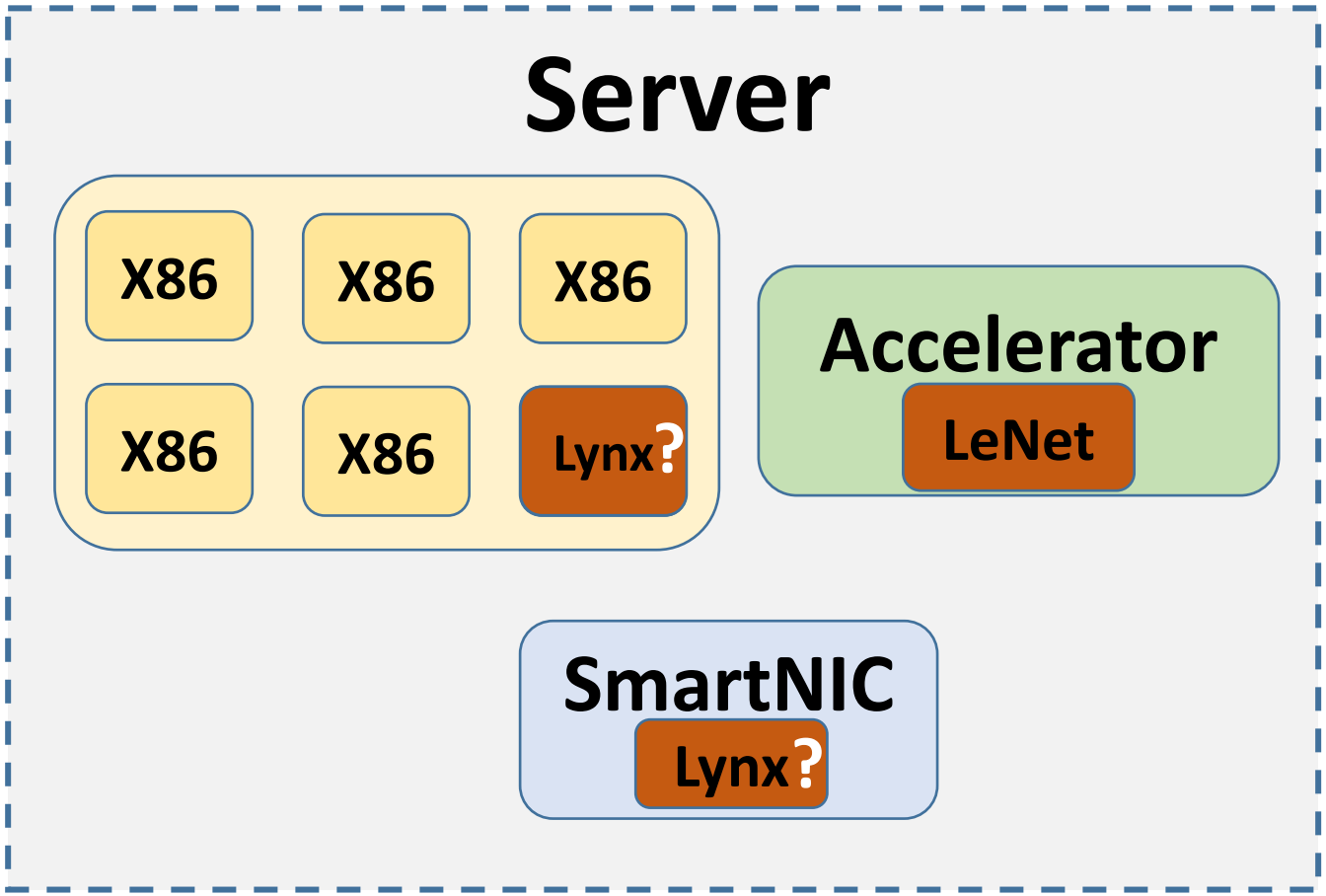


Applications

memcached

What is the best mapping?

Why not using the SmartNIC as another CPU?

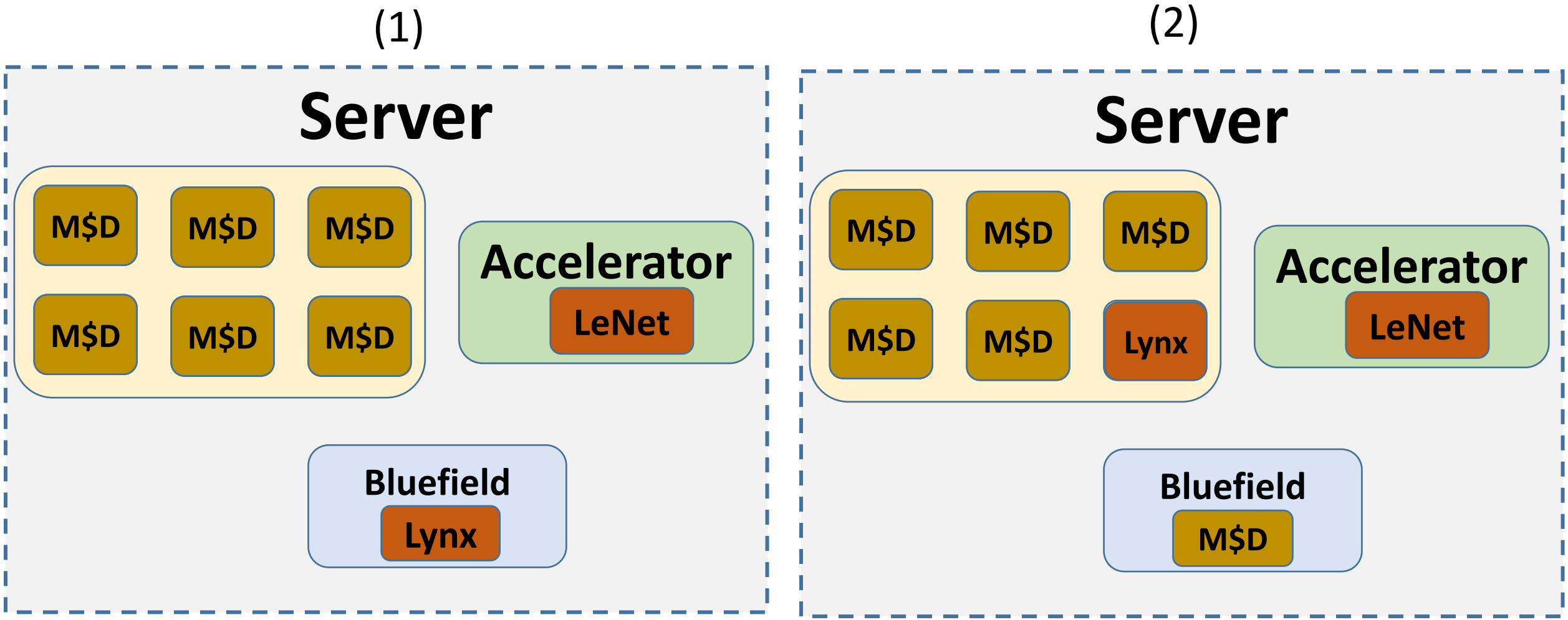


Applications

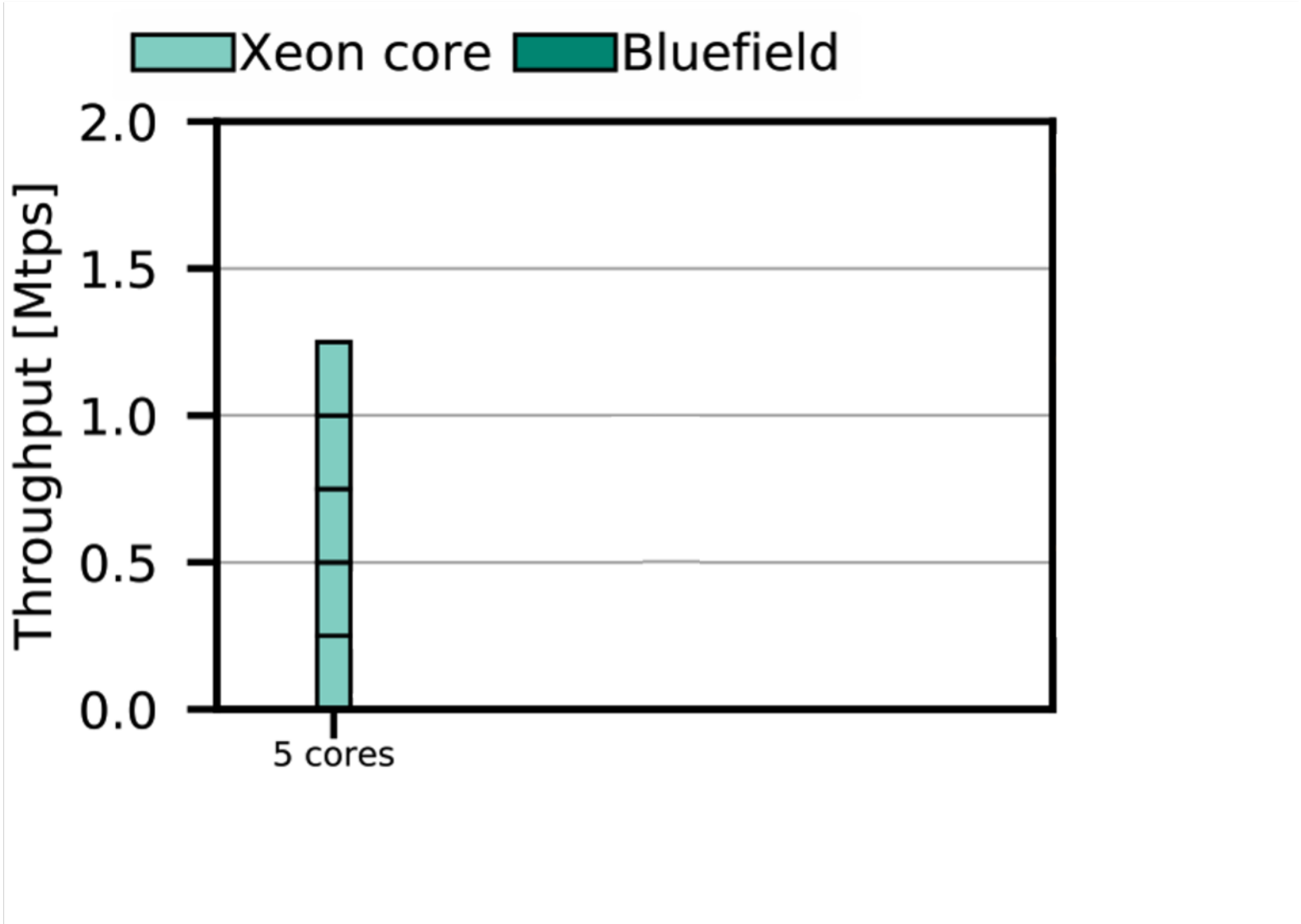
memcached

What is the best mapping?

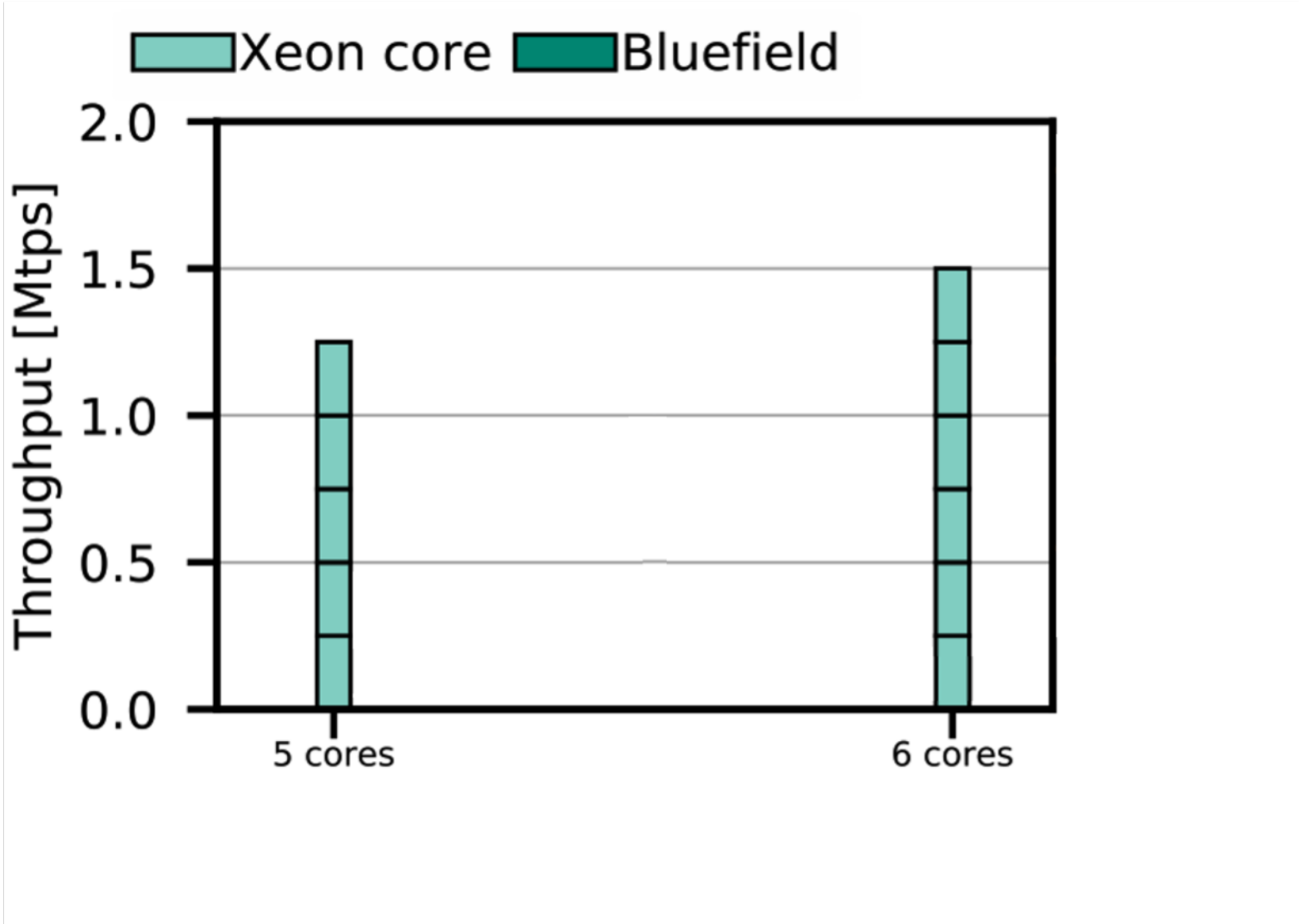
Where should we run Memcached?



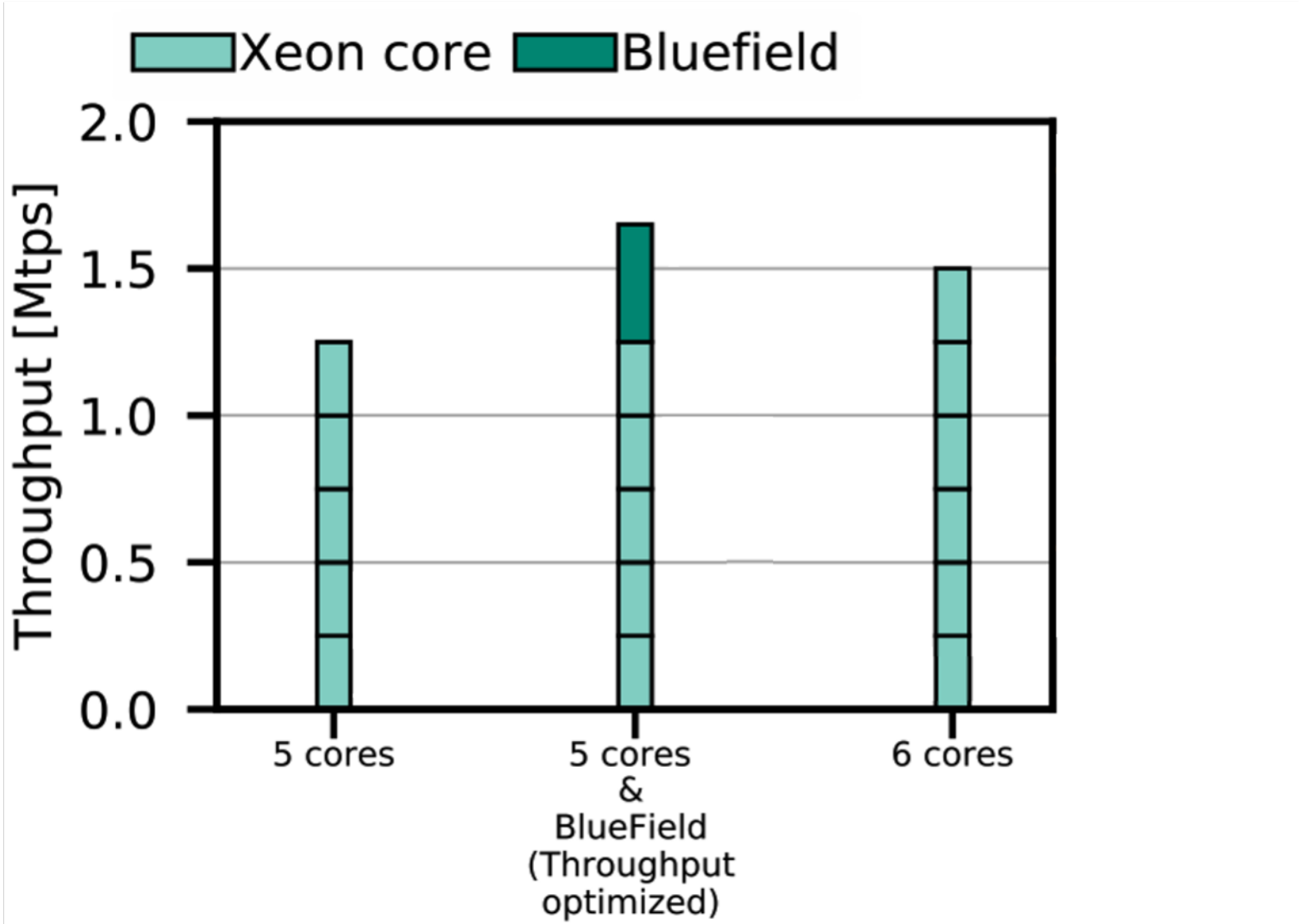
Memcached Performance



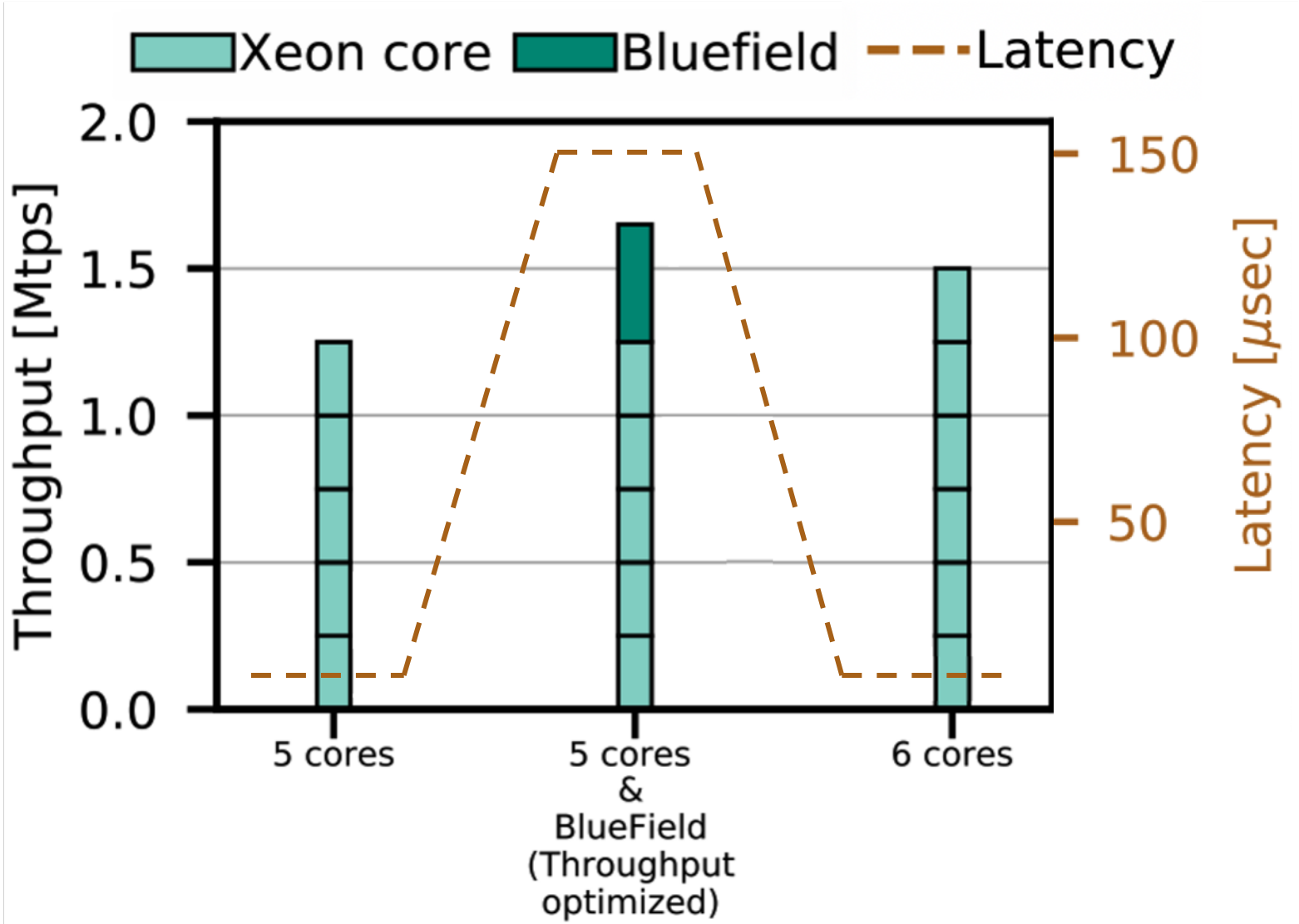
Memcached Performance



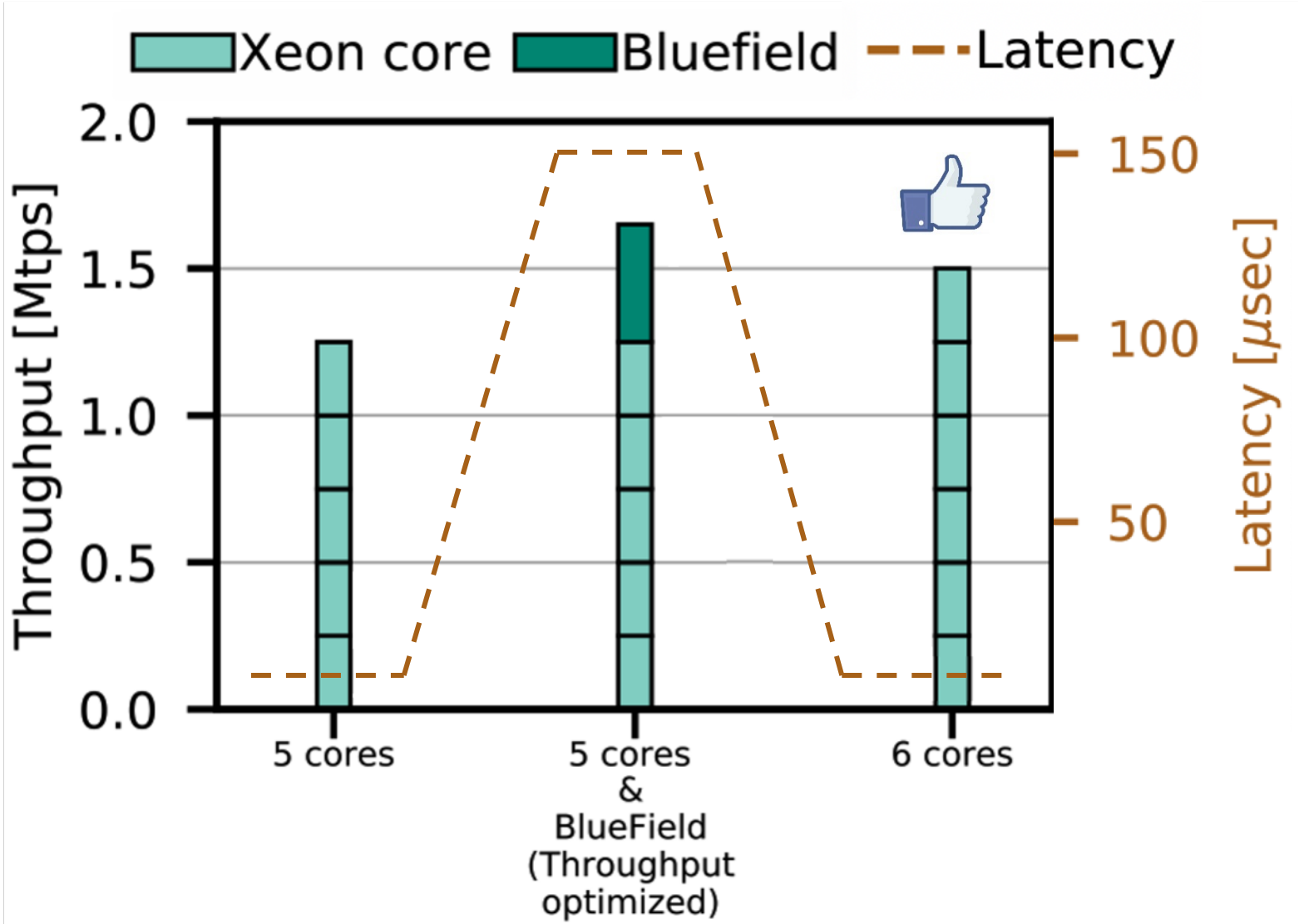
Memcached Performance



Memcached Performance



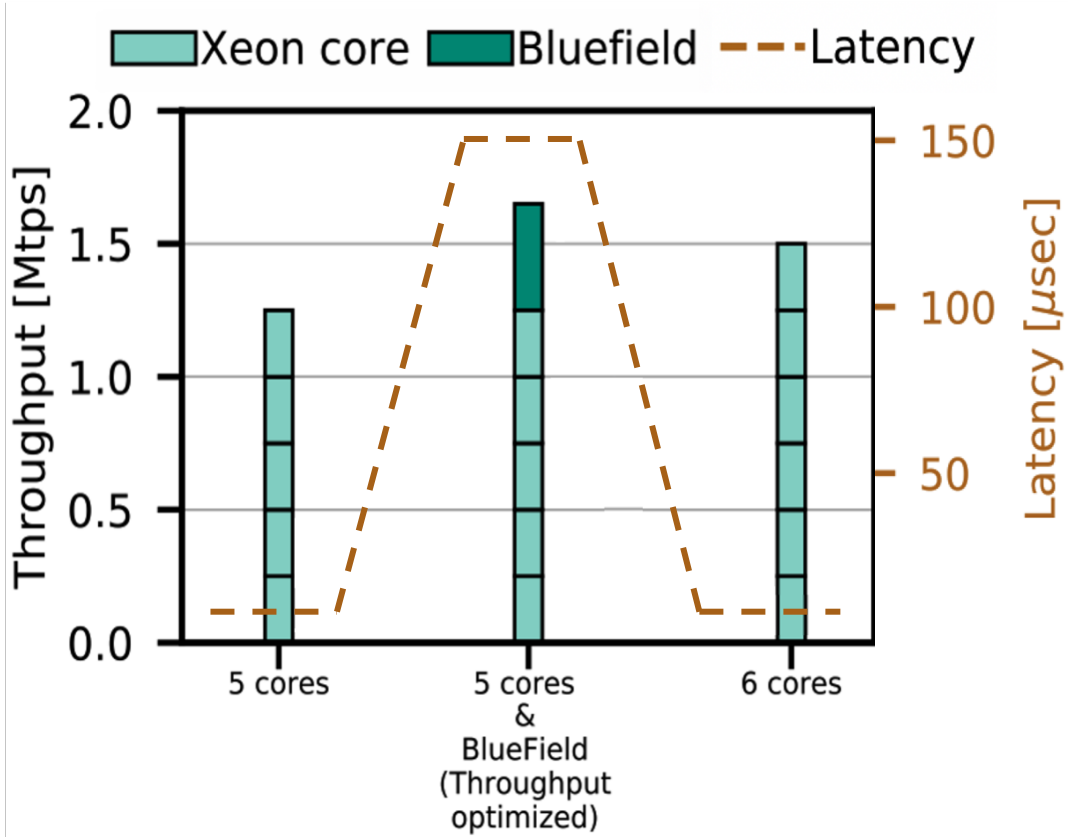
Memcached Performance



Memcached Performance

Takeaway

Lynx on Bluefield may achieve higher system efficiency compared to using Bluefield for standard server workloads.



Agenda

- Motivation and Background
- Design
- Evaluation
- Conclusions

Conclusions

- Accelerated servers can run with 0% CPU utilization

- SmartNICs provide a unique opportunity to manage accelerators and run generic network servers

Thanks!

Questions?