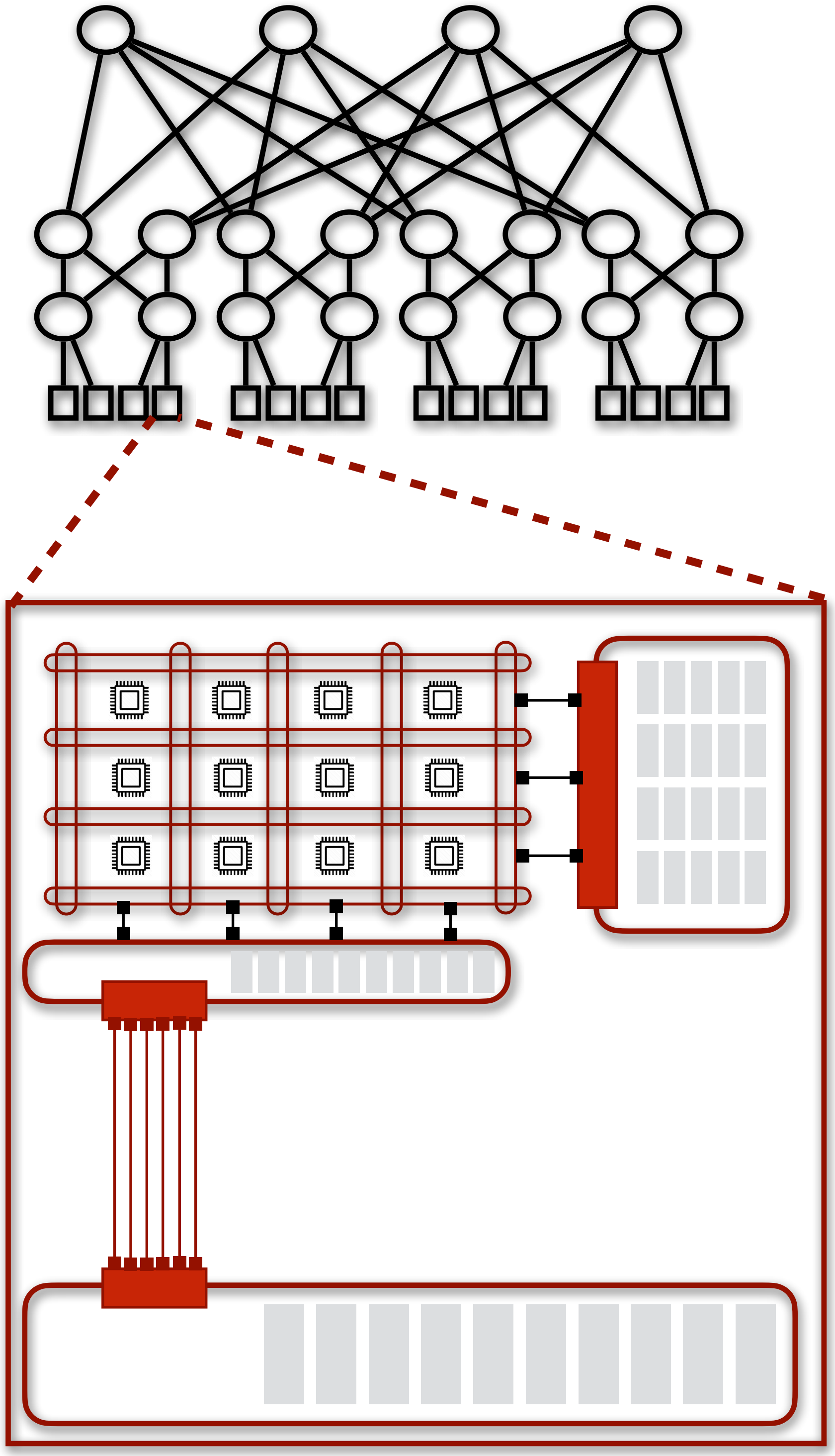
Understanding Host Interconnect Congestion



Saksham Agarwal

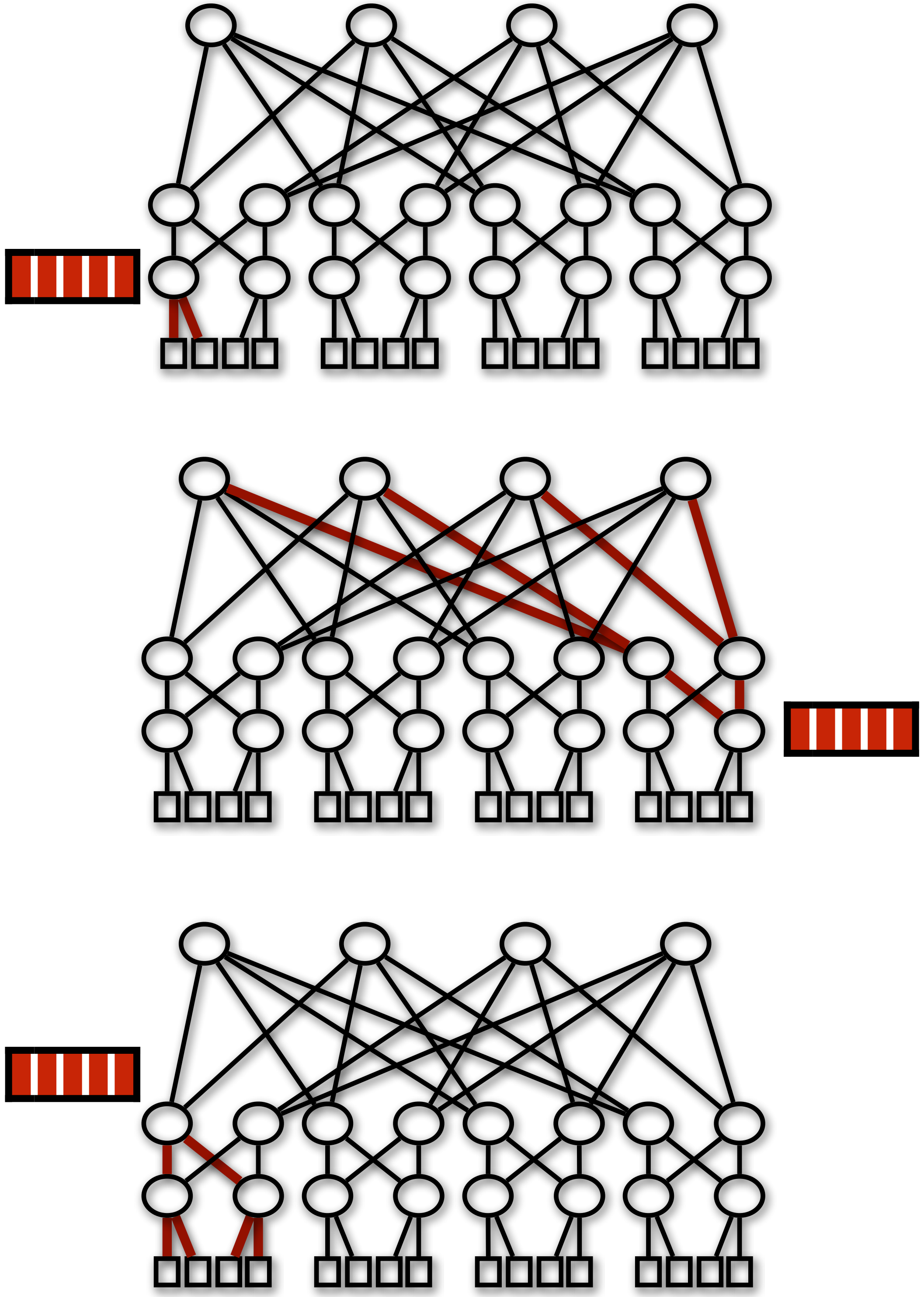
Cornell University

In collaboration with:
Rachit Agarwal (Cornell)
Behnam Montazeri (Google)
Masoud Moshref (Google)
Khaled Elmeleegy (Google)
Luigi Rizzo (Google)
Marc Asher de Kruijf (Google)
Gautam Kumar (Google)
Sylvia Ratnasamy (Google & UC Berkeley)
David Culler (Google)
Amin Vahdat (Google)

Conventional wisdom: Congestion in the network core

Congestion happens in the network core: at switches

Due to oversubscribed topologies, incast traffic pattern, and/or poor load balancing



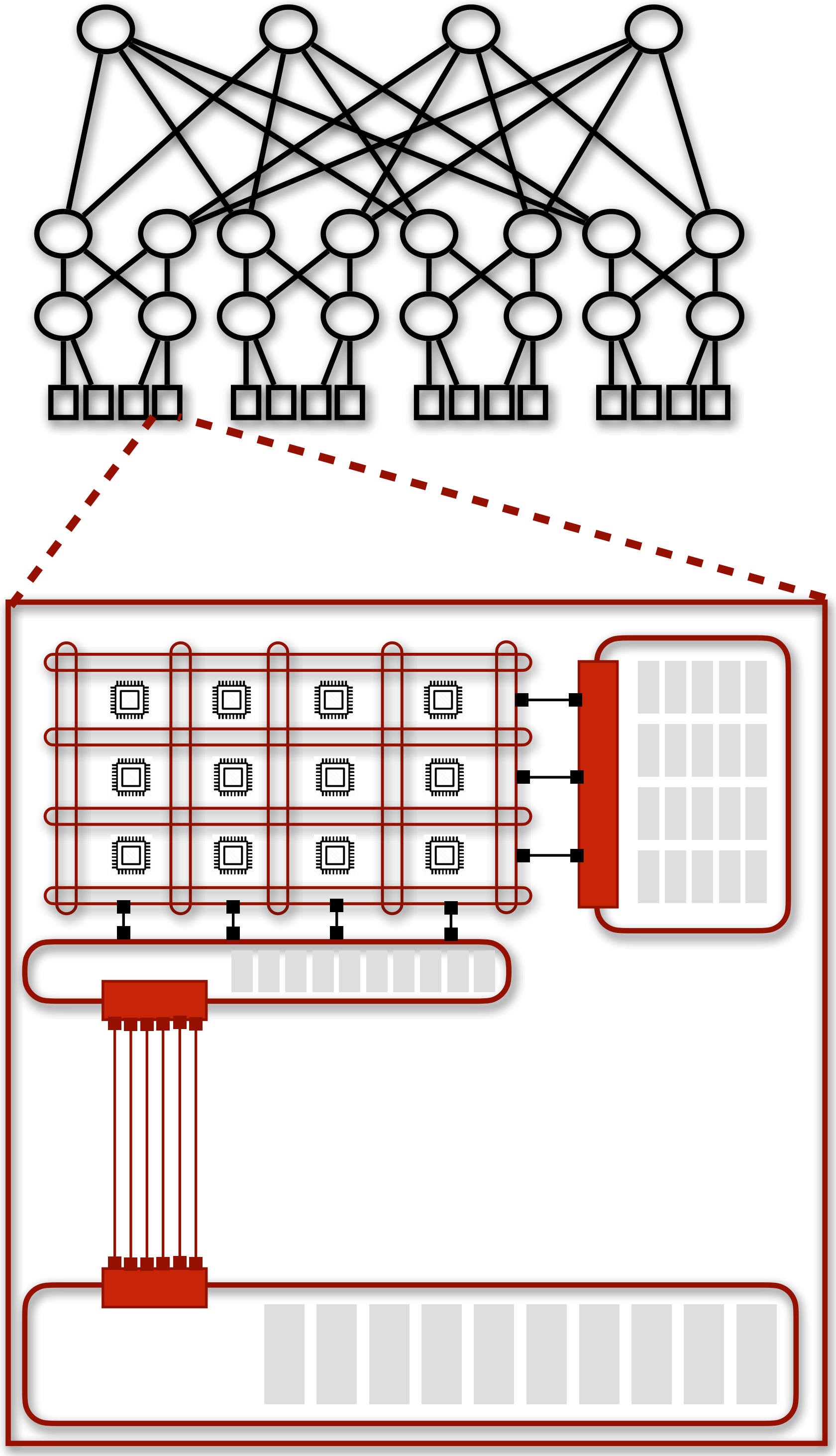
Decades of work; deep understanding of:

- Reasons for congestion
- Congestion signals
- Congestion response
- …..

This work: Host Congestion

Due to emergence of host interconnect bottlenecks

Data path between the NIC and the CPU/memory



○ **Understanding host congestion**

And its impact

○ **Root causes of host congestion**

Building a deeper understanding

○ **Towards resolving host congestion**

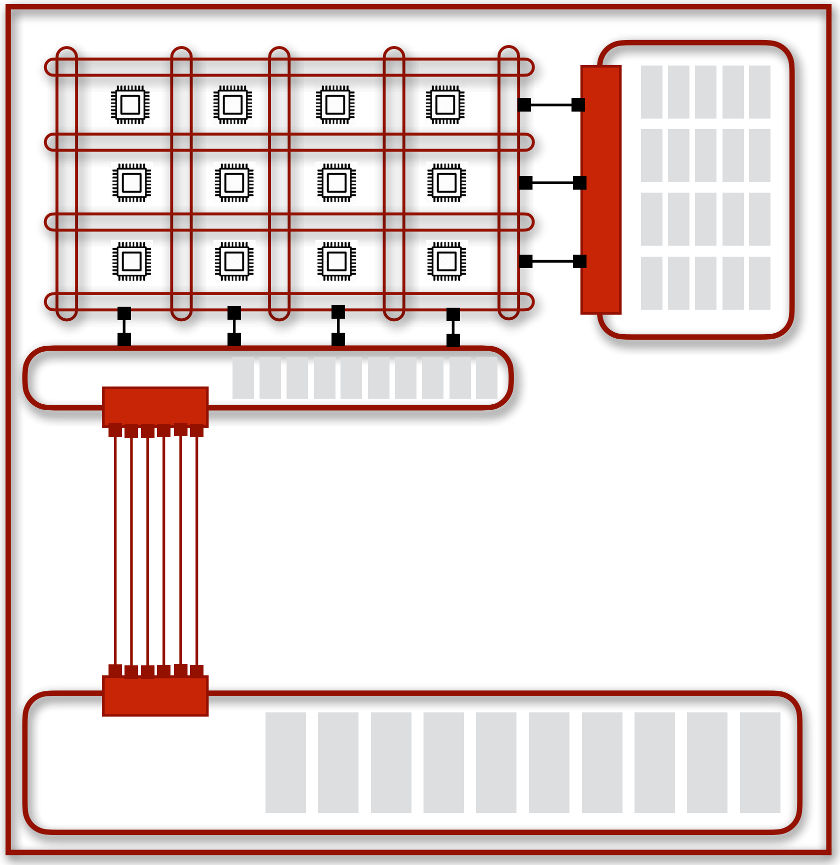
Need for:

- New host architectures
- New congestion signals
- New congestion response

This work: Host Congestion

Due to emergence of host interconnect bottlenecks

Data path between the NIC and the CPU/memory



○ **Understanding host congestion**

And its impact

○ **Root causes of host congestion**

Building a deeper understanding

○ **Towards resolving host congestion**

Need for:

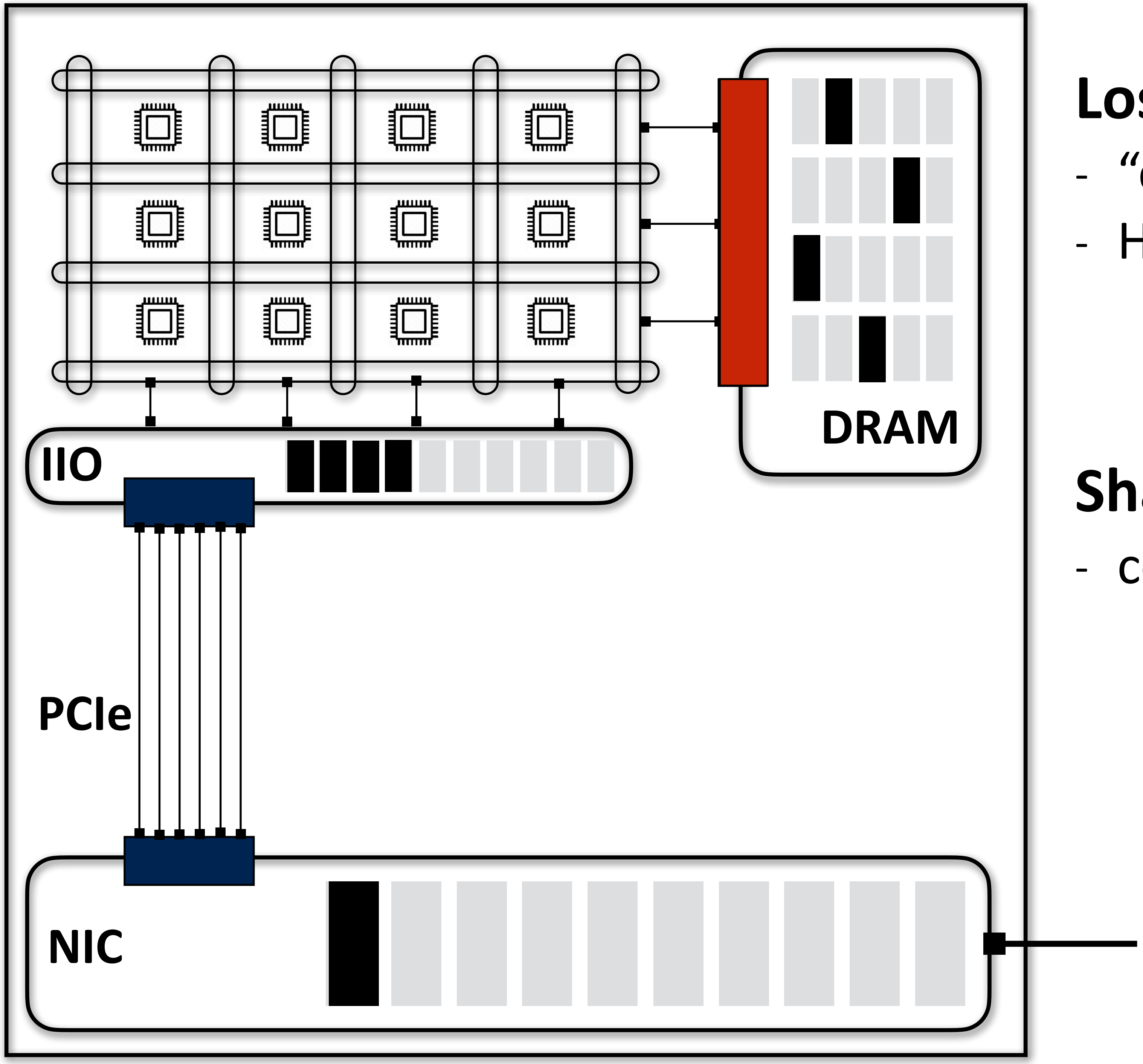
- New host architectures
- New congestion signals
- New congestion response

Host Interconnect: a brief primer

Host interconnect comprises multiple subsystems

Peripheral interconnect (PCIe), processor interconnect, memory channels, etc.

All operating independently in a closed-loop system (to enable losslessness)



Lossless interconnect

- “credit”-based
- Hop-by-hop

Shared interconnect

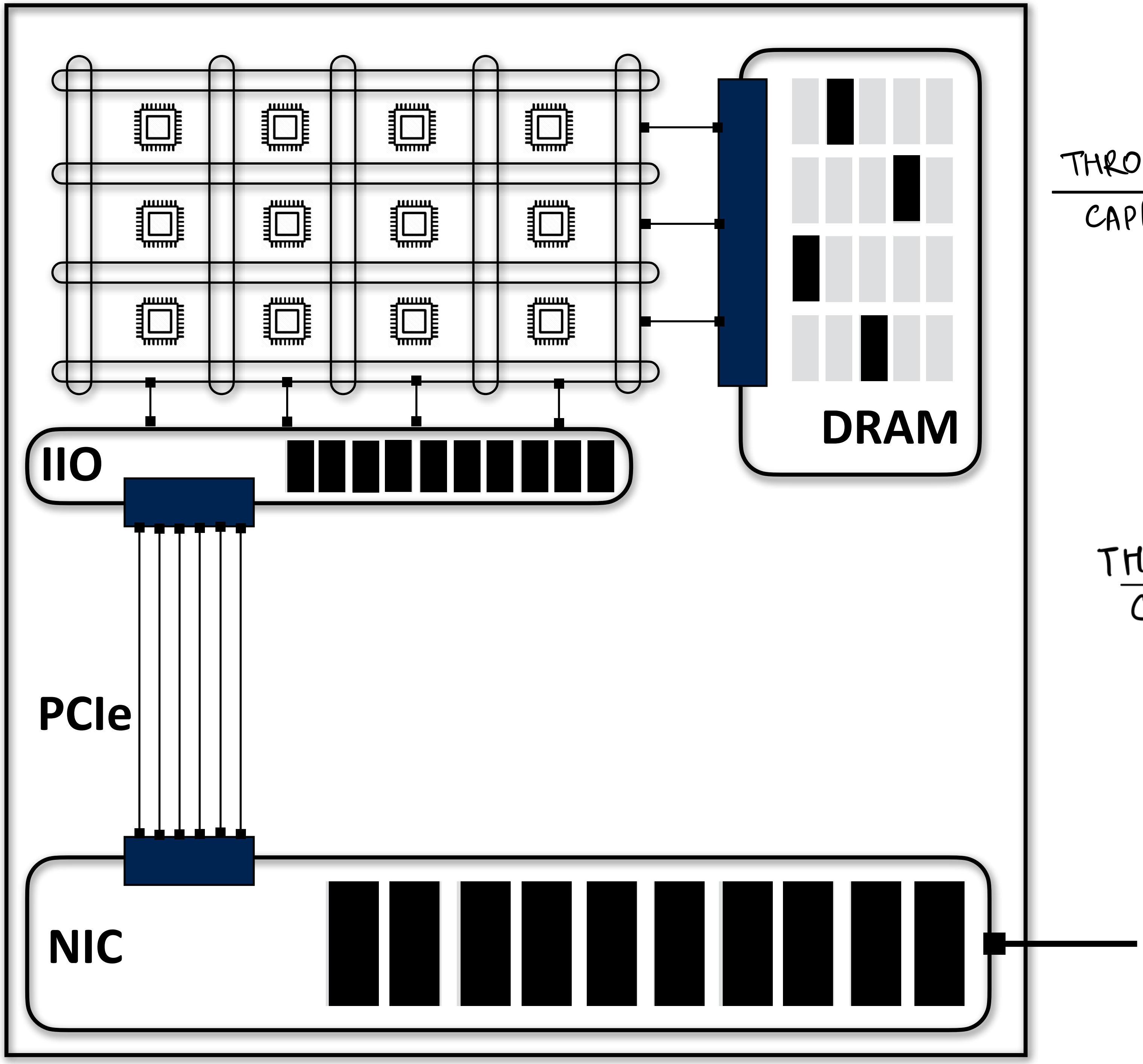
- compute & peripheral traffic share:
  - Both processor interconnect
  - And, memory channels

Host Congestion

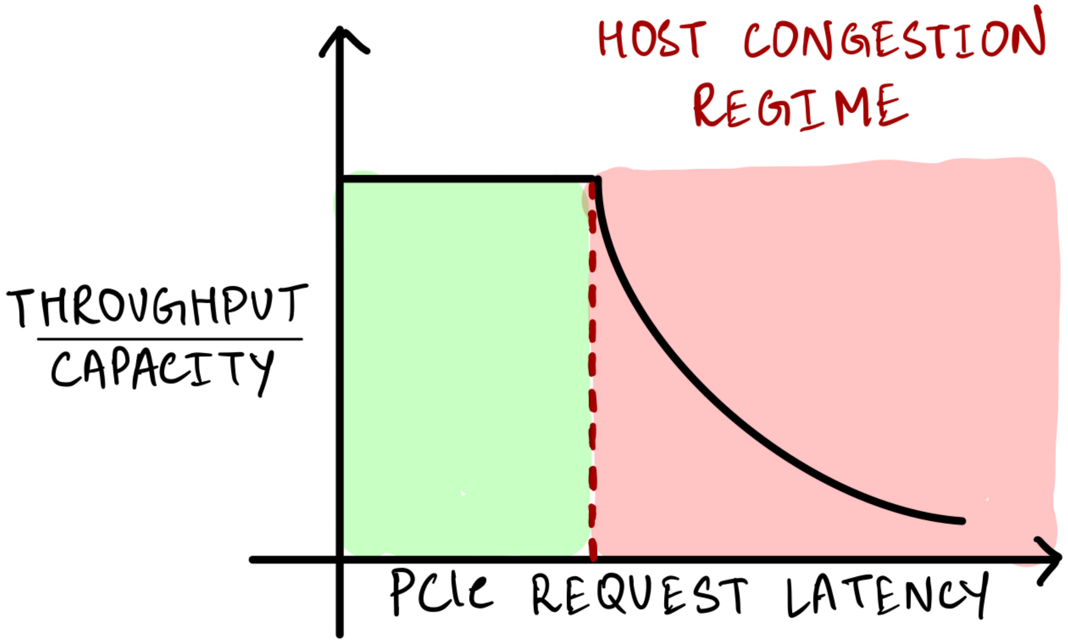
NIC unable to drain packets at the same rate at which it receives packets

PCIe bandwidth is underutilized

NIC buffers build up even before senders can respond; packets dropped



$$\frac{\text{THROUGHPUT}}{\text{CAPACITY}} = \text{MIN}\left\{1, \frac{\text{\# MAX CREDITS} \times \text{REQ SIZE}}{\text{PCIe REQUEST LATENCY}}\right\}$$

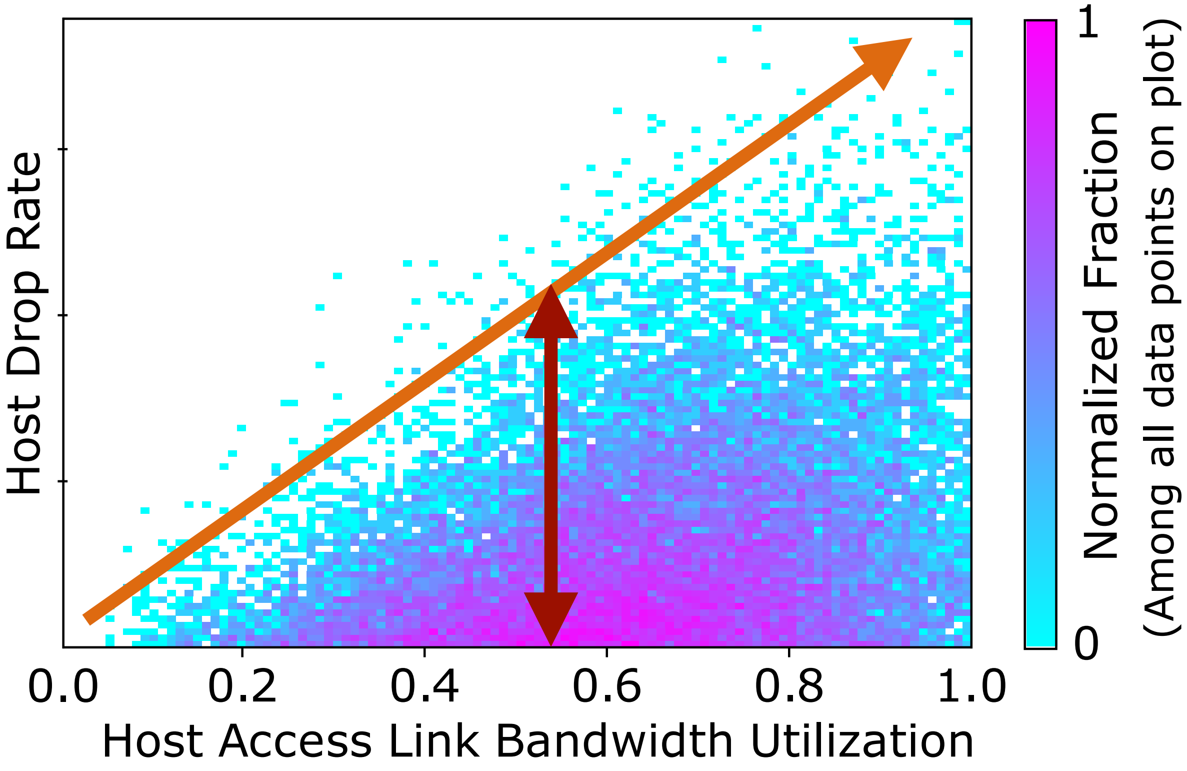


Host Congestion in production clusters

Google production cluster

Runs SNAP with Swift as congestion control protocol (also Linux + TCP)

Minimal in-network congestion, and auto-scaling for CPU bottlenecks



As access link utilization increases

=> more drops

Even when access link far from saturated

=> significant drop rate

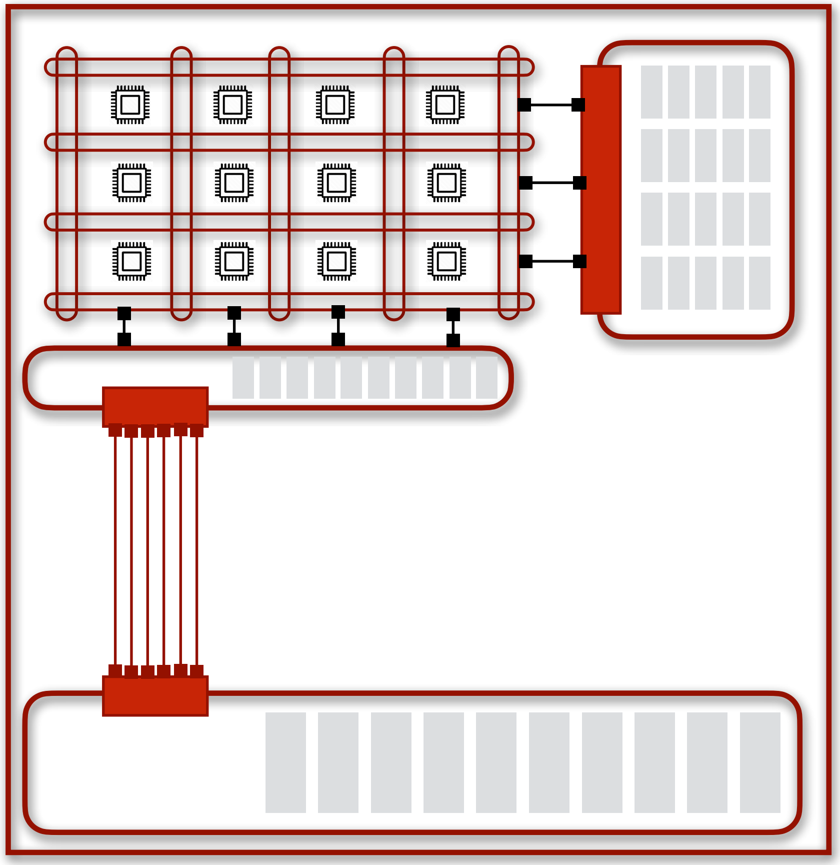
Impact of host congestion

Poor isolation, inflated tail latency, low throughput

This work: Host Congestion

Due to emergence of host interconnect bottlenecks

Data path between the NIC and the CPU/memory



○ **Understanding host congestion**

And its impact

○ **Root causes of host congestion**

Building a deeper understanding

○ **Towards resolving host congestion**

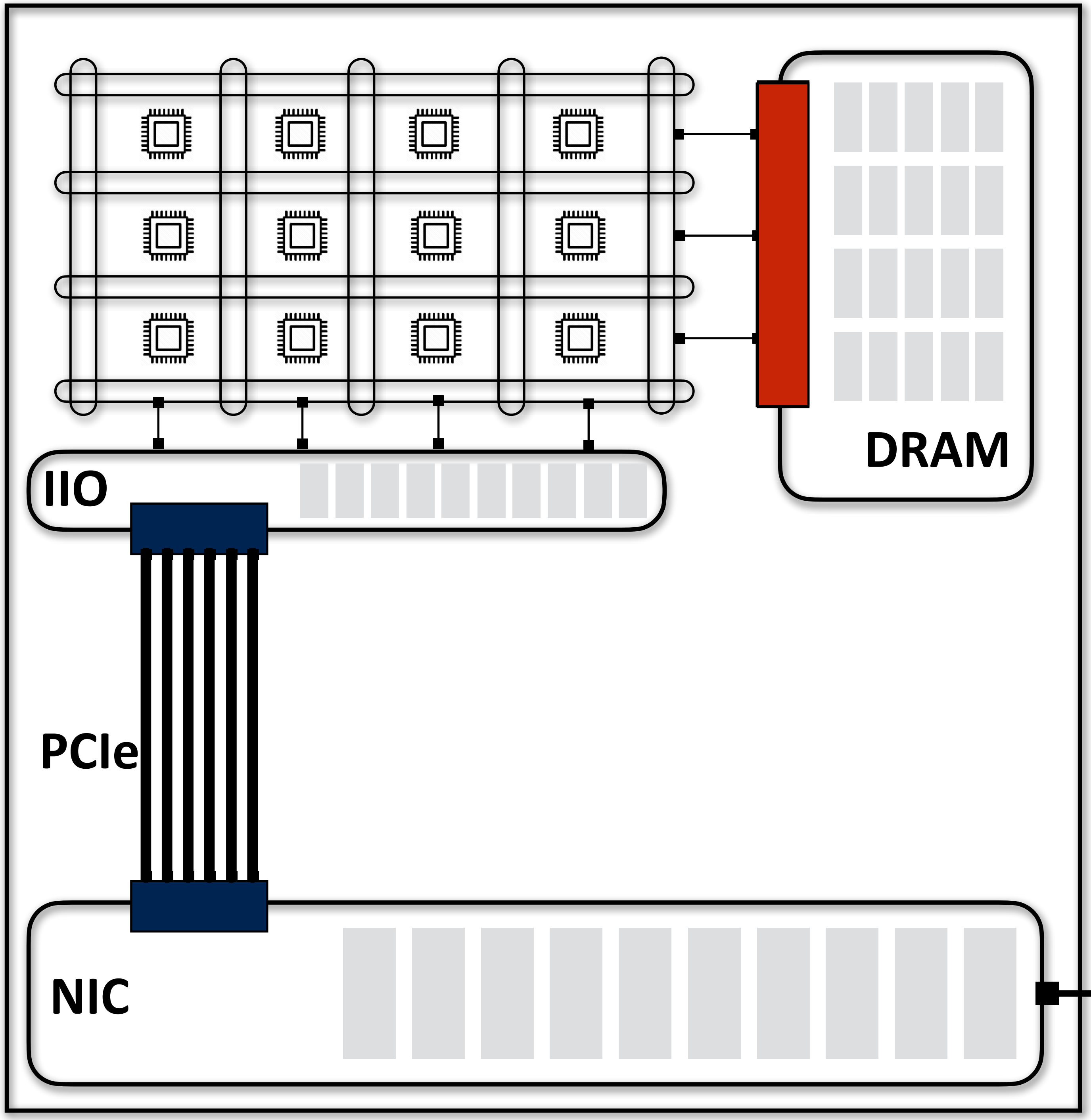
Need for:

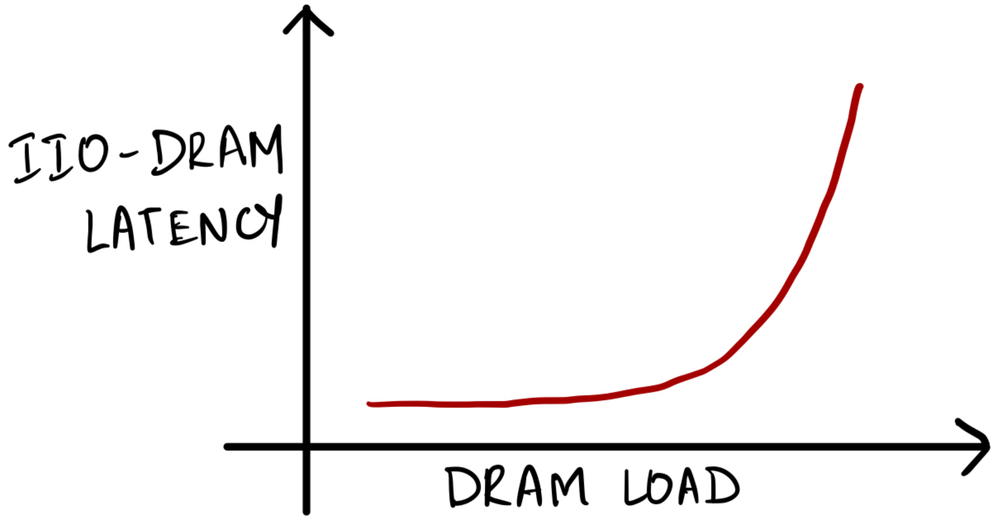
- New host architectures
- New congestion signals
- New congestion response

Host Congestion due to Host Interconnect Bottlenecks [1]

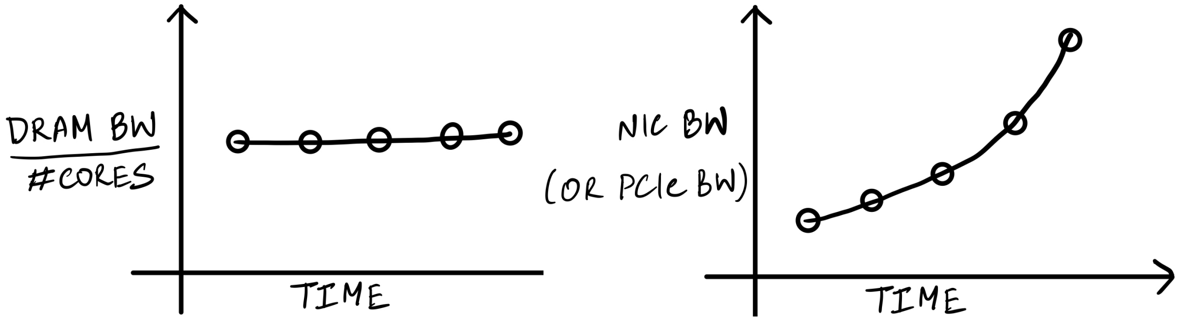
Reducing ratio of DRAM bandwidth to IO bandwidth (+CPU bandwidth)

+ Poor isolation at the DRAM controller





PROBLEM GOING TO GET WORSE

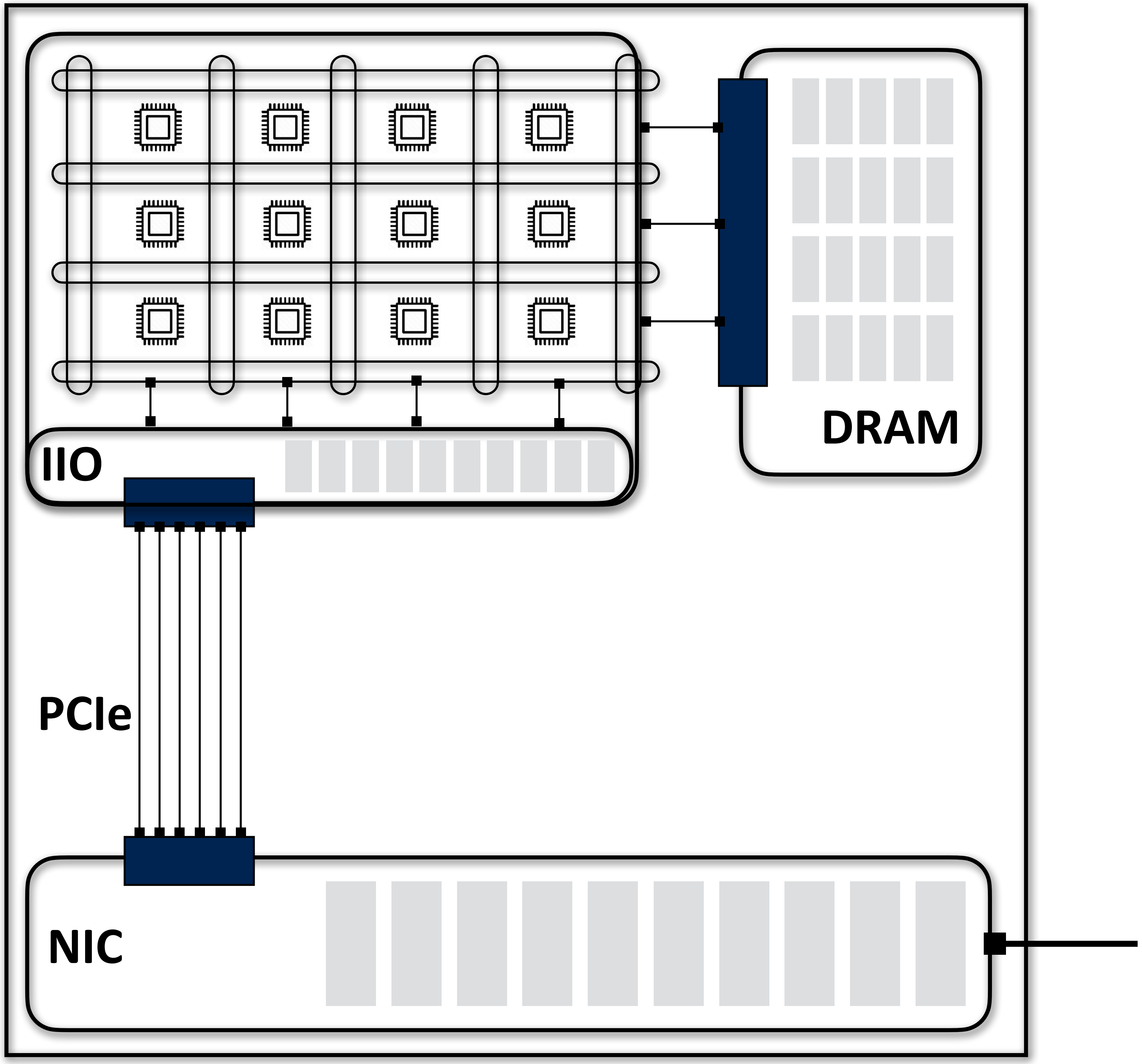


Host Congestion due to Host Interconnect Bottlenecks [2]

Inefficient mechanisms for memory protection

NIC deals with virtual addresses; final operations on physical addresses

IOMMU translates addresses using an IO page table; IOTLB is cache for IO page table

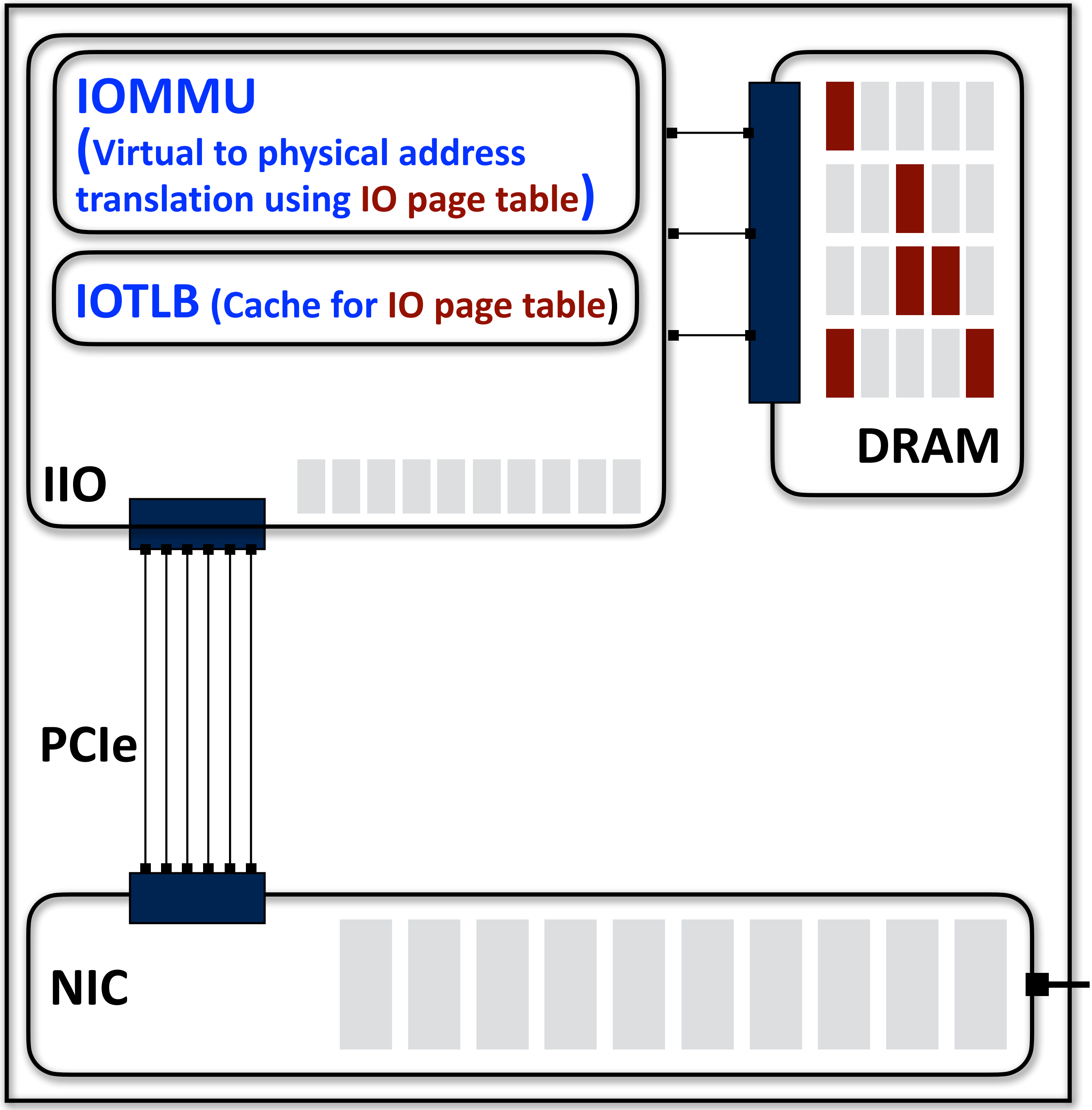


Host Congestion due to Host Interconnect Bottlenecks [2]

Inefficient mechanisms for memory protection

NIC deals with virtual addresses; final operations on physical addresses

IOMMU translates addresses using an IO page table; IOTLB is cache for IO page table



IOMMU

(Virtual to physical address translation using IO page table)

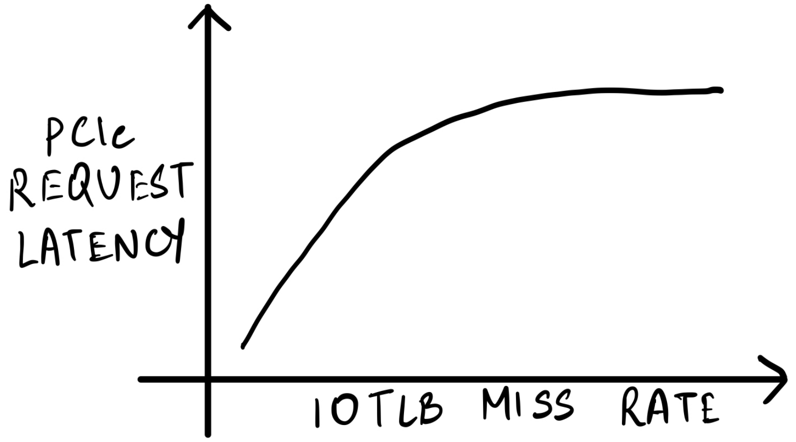
IOTLB (Cache for IO page table)

IIO

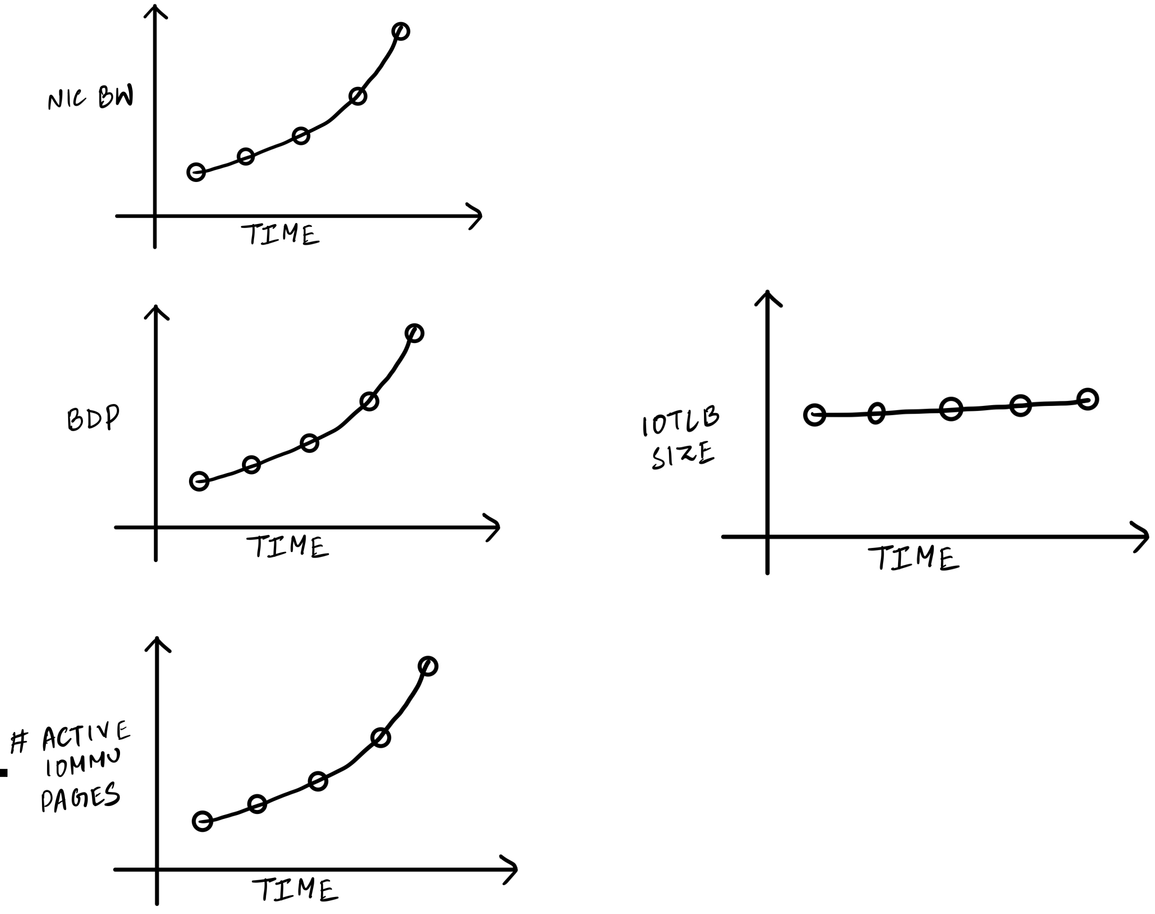
DRAM

PCIe

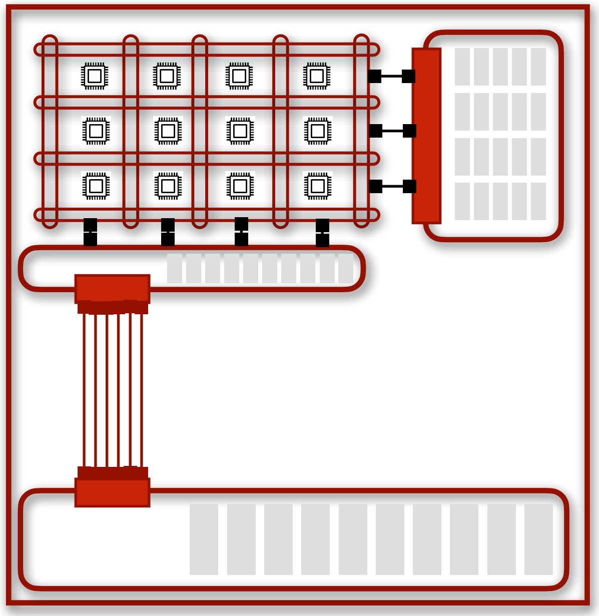
NIC



PROBLEM GOING TO GET WORSE



Host Congestion: more details in the paper

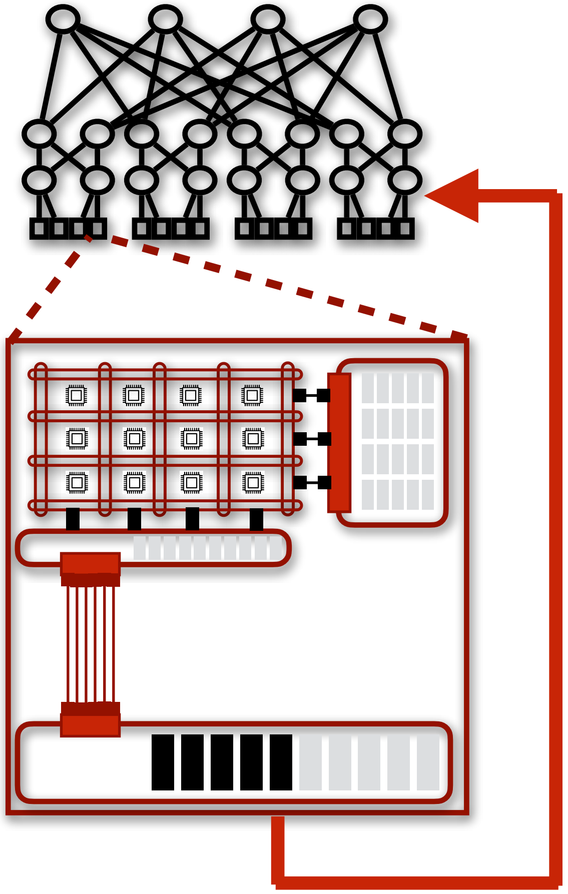


Workloads that lead to host congestion

Common workloads: one-to-one, incast, all-to-all

Observed in large-scale Google production clusters

- Results reproducible on commodity machines with Linux
- Paper: minimalistic workloads for reproducing results
- Reach out to me for help.



Existing CC protocols do not account for host congestion

Reducing rate =/=> reduce contention (e.g., IOMMU)

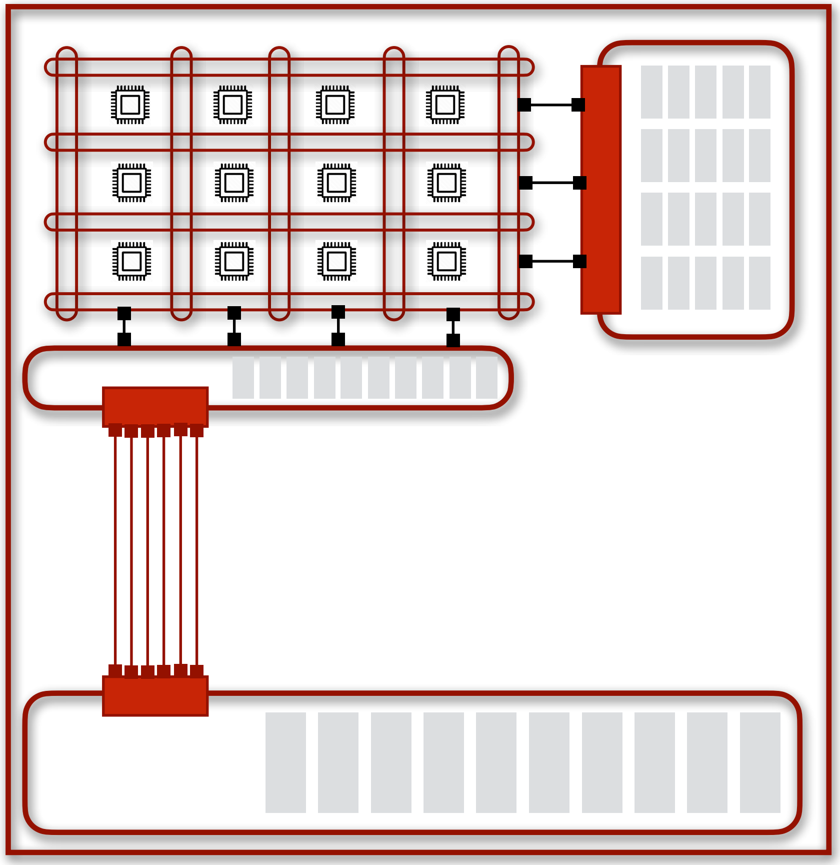
Several unexpected behavior

- Non-monotonic relationship between contention & drops
- Using Hugepages results in higher drops
- …

This work: Host Congestion

Due to emergence of host interconnect bottlenecks

Data path between the NIC and the CPU/memory



○ **Understanding host congestion**

And its impact

○ **Root causes of host congestion**

Building a deeper understanding

○ **Towards resolving host congestion**

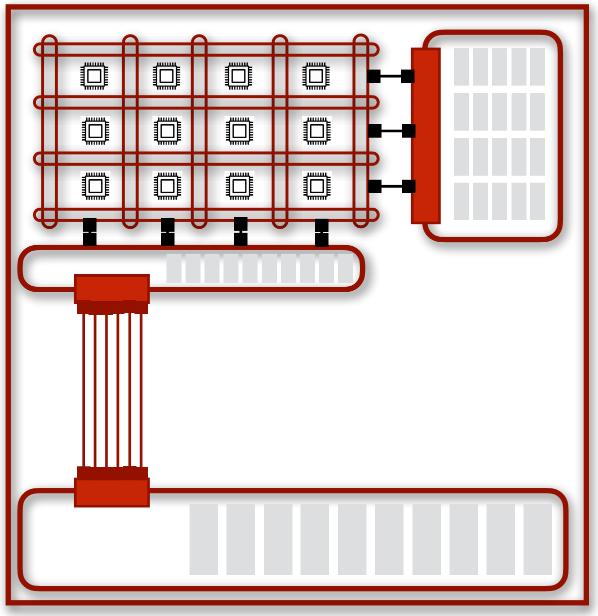
Need for:

- New host architectures
- New congestion signals
- New congestion response

Host Congestion: Looking forward

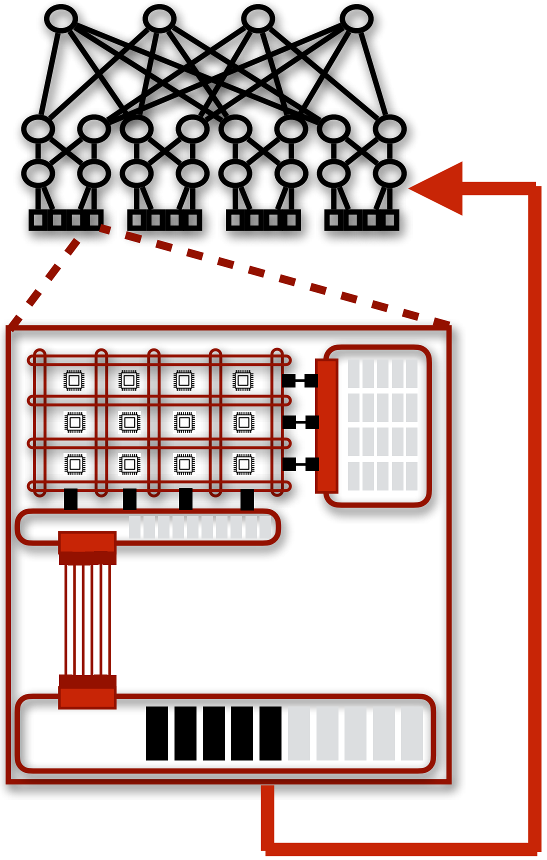
Need to rethink host architecture, network stack, network protocols

Bring together ideas from networking, operating systems, and architecture



Rethink Host architecture

- PCIe enhancements (e.g., CXL)
  - Stronger semantics, lower latency
- Memory protection mechanisms (e.g., ATS)
  - Address translation offload
- Memory controller architecture
  - Sharing mechanisms for memory channels



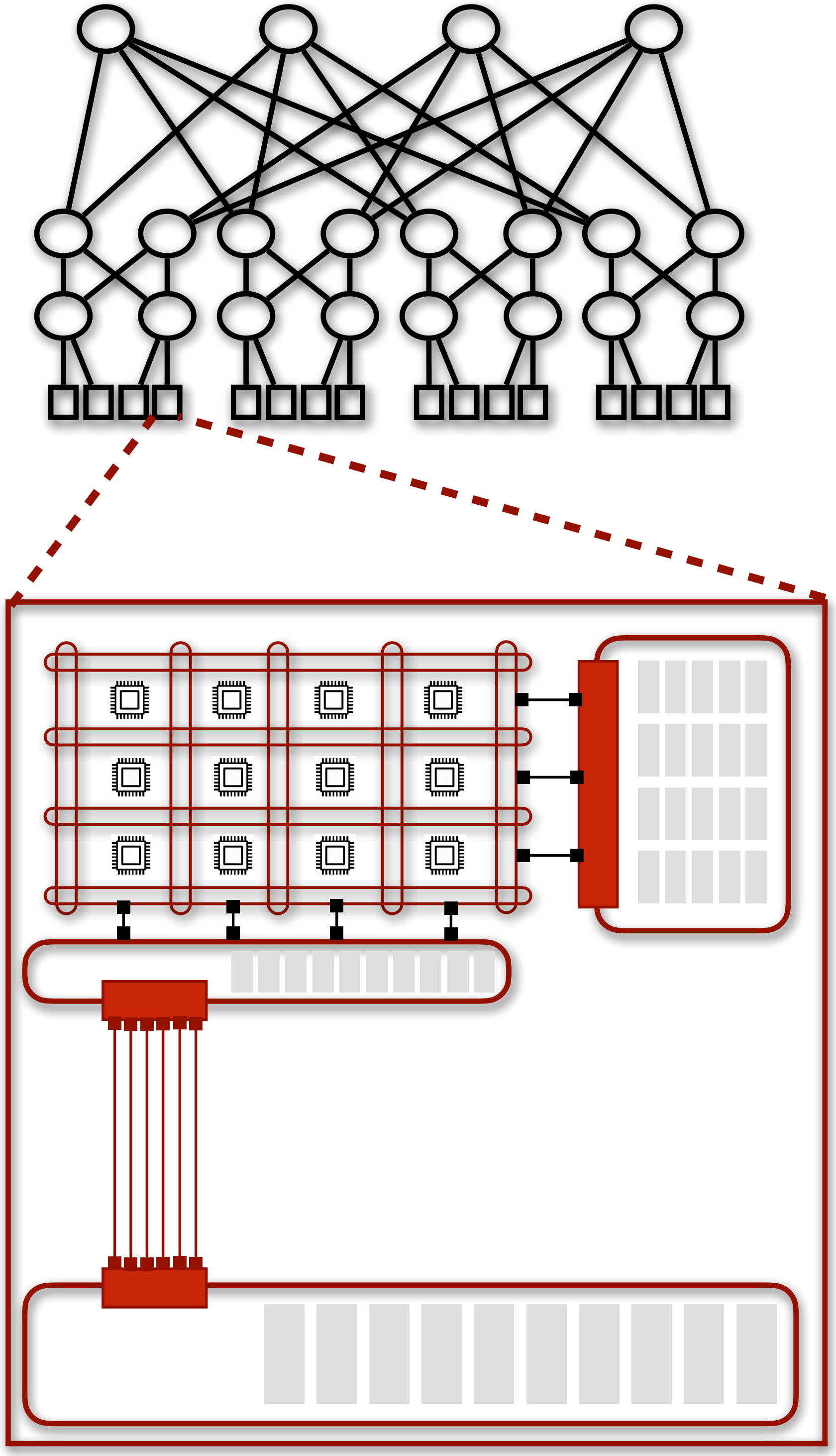
Rethink network stacks and protocols

- New congestion signals
  - from “outside” the network
  - e.g., memory load, fragmentation, etc.
- New congestion response
  - Different for different root causes (memory vs IOMMU)?
  - sub-RTT response

Host Congestion

Due to emergence of host interconnect bottlenecks

Data path between the NIC and the CPU/memory



○ **Understanding host congestion**

And its impact

○ **Root causes of host congestion**

Building a deeper understanding

○ **Resolving host congestion**

Need for:

- New host architectures
- New congestion signals
- New congestion response