SC23
Denver, CO | i am hpc.

RISA: Round-Robin Intra-Rack Friendly Scheduling Algorithm for Disaggregated Datacenters

Rashadul Kabir, Colorado State University, Fort Collins, CO, USA

Ryan G. Kim, Intel Labs, Hillsboro, OR, USA

Mahdi Nikdast, Colorado State University, Fort Collins, CO, USA

Electronic–PhotoniC System Design (ECSyD) Laboratory

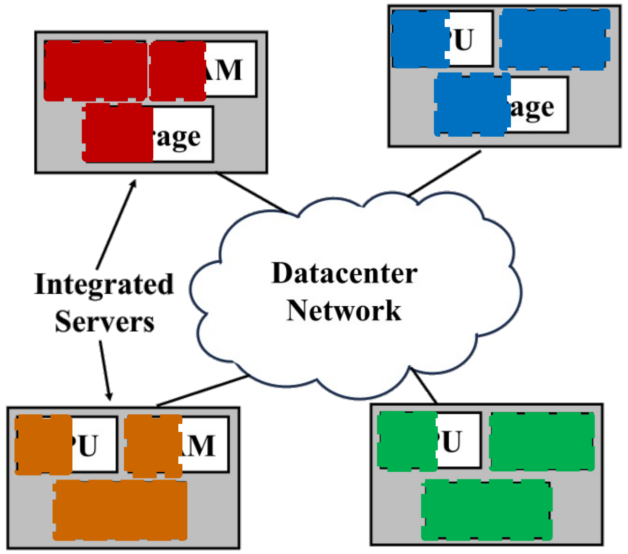
Outline

- Introduction
- Disaggregated datacenter and related work
- Optical switch model
- RISA: Round-Robin Intra-Rack Friendly Scheduling Algorithm for Disaggregated Datacenters
- Discussion of simulation results
- Conclusion

Introduction: Why Disaggregate?

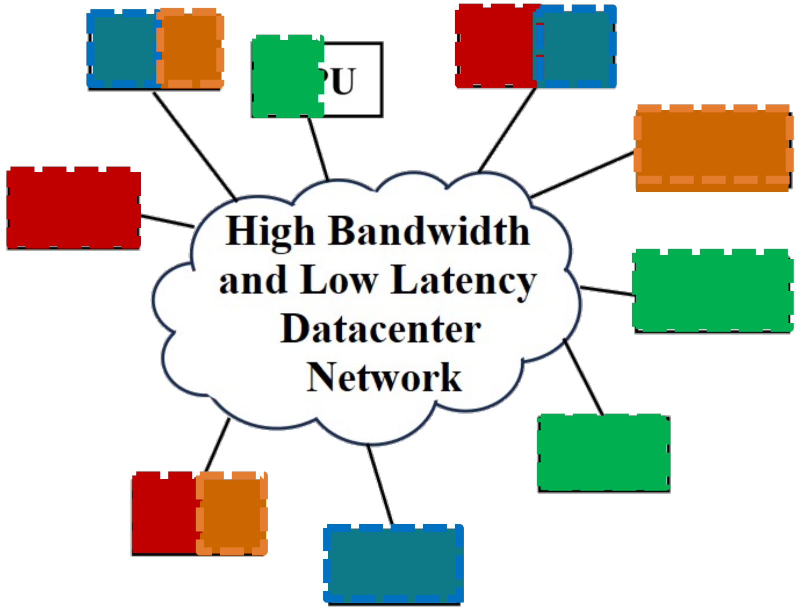
- Modern applications have varying compute requirements, e.g.
  - CPU intensive (requires more CPU)
  - RAM intensive

- Traditional datacenter
  - Fixed resource configurations
  - Partial compute resource utilization



Traditional Datacenter

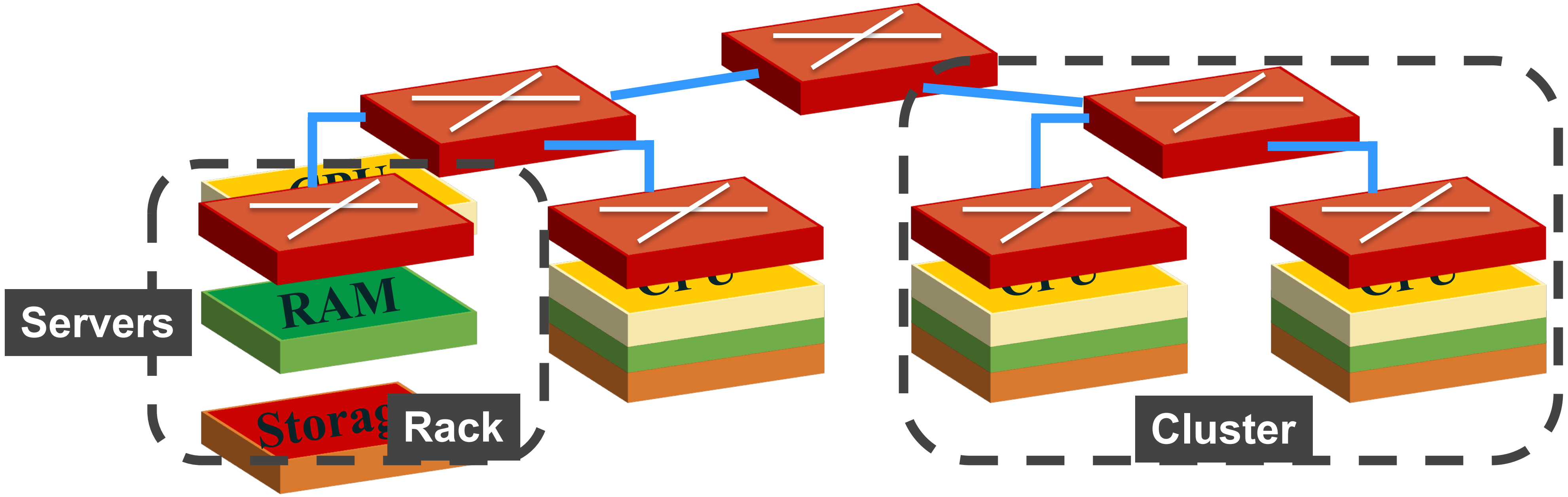
- Solution: Disaggregated Datacenter (DDC)
  - Requires fewer compute resources
  - High compute resource utilization



Disaggregated Datacenter

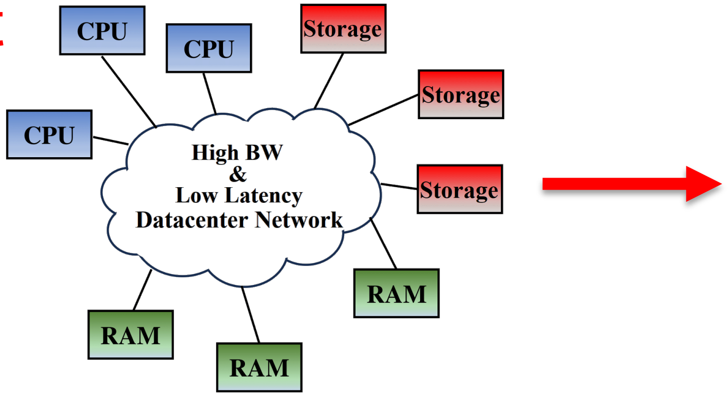
Challenges: How can this work?

- ➢ Disaggregated datacenters arranged in **servers**, **racks**, and **clusters**



How do we schedule VMs?

- ➢ Network infrastructure to support DDC is expensive!
  - ▪ Capital cost
  - ▪ Energy

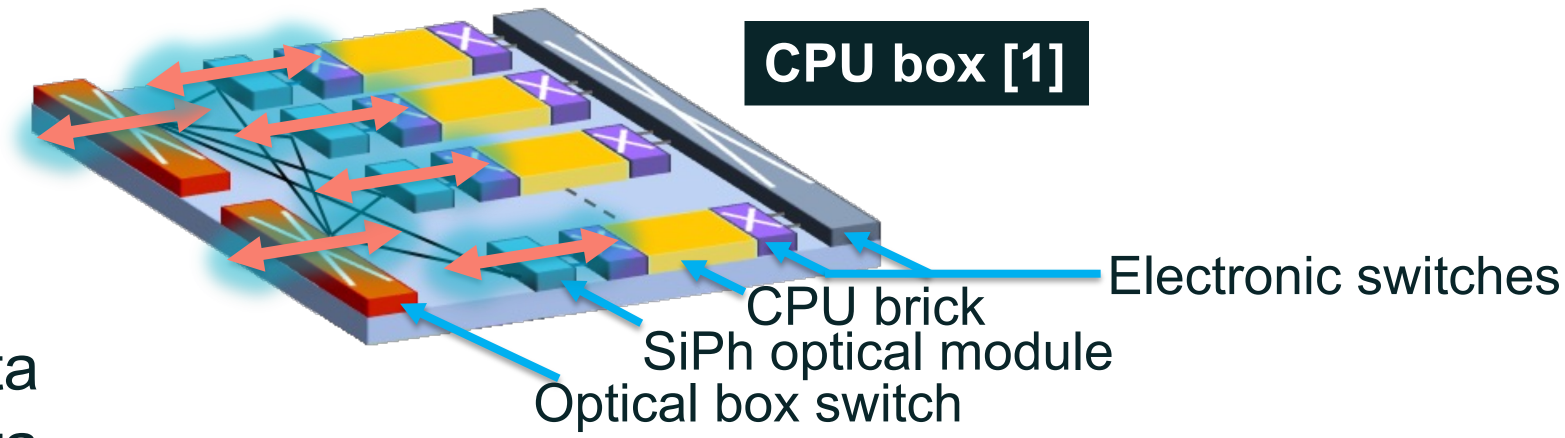


Goals

- **Well-coordinated scheduling of CPU, RAM, storage, and network**
  - High compute resource utilization (same as state-of-the-art)
  - **Low network utilization**
    - Low power consumption
    - Low CPU-RAM round-trip latency
  - Low-cost scheduling policy

Disaggregated Datacenter (DDC)

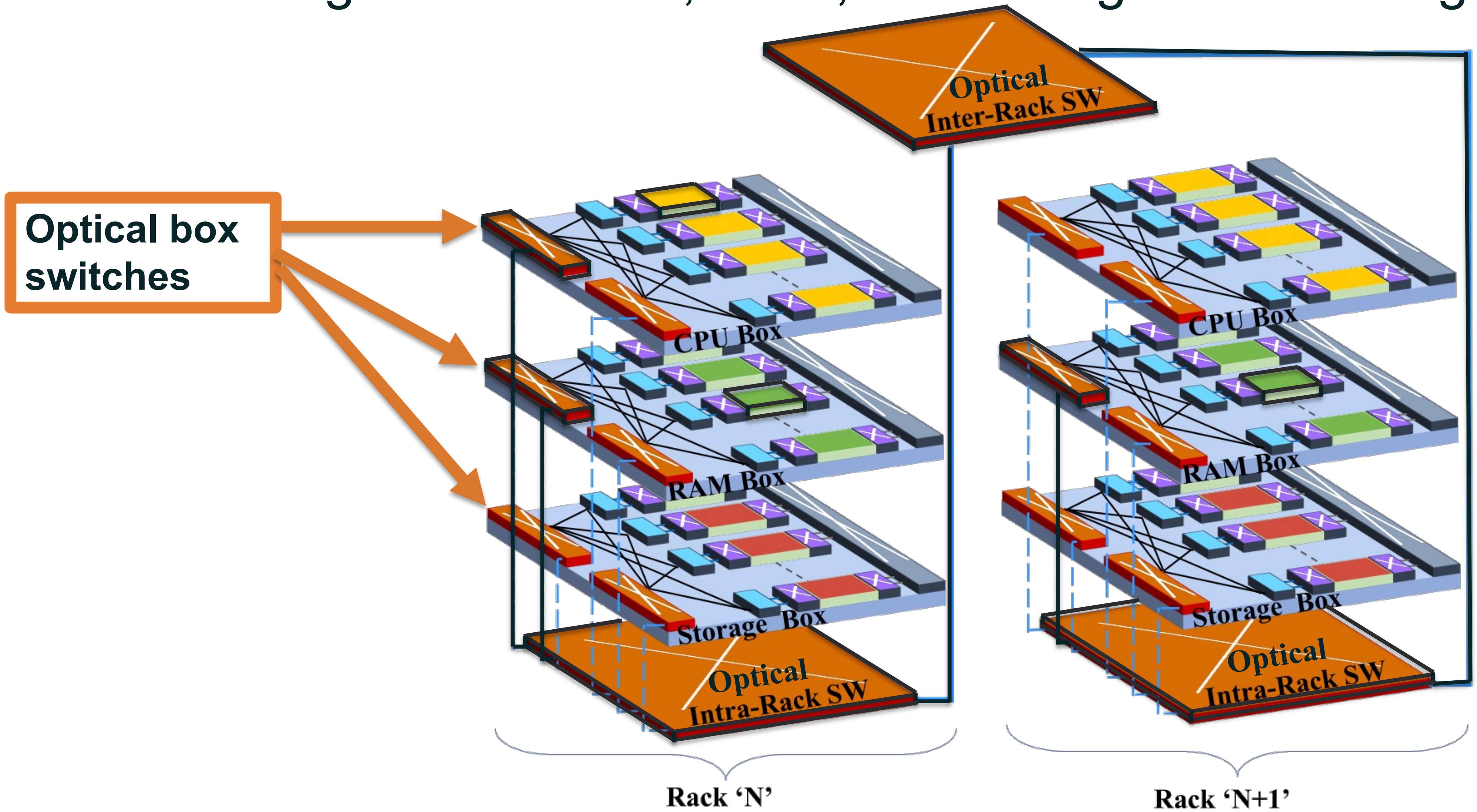
- One compute resource per server (box)
  - CPU brick: 64 cores
    - Electronic switches allow
      - Intra-brick communication
      - Inter-brick communication
  - SiPh Optical module
    1. Electronic data – Optical data
    2. Optical data – Electronic data
  - Optical box switch
    - Communication with optical intra-rack switch



[1] G. Zervas, H. Yuan, A. Saljoghei, Q. Chen, and V. Mishra, "Optically disaggregated data centers with minimal remote memory latency: Technologies, architectures, and resource allocation [Invited]," Journal of Optical Communications and Networking, 2018.

DDC used as Case Study

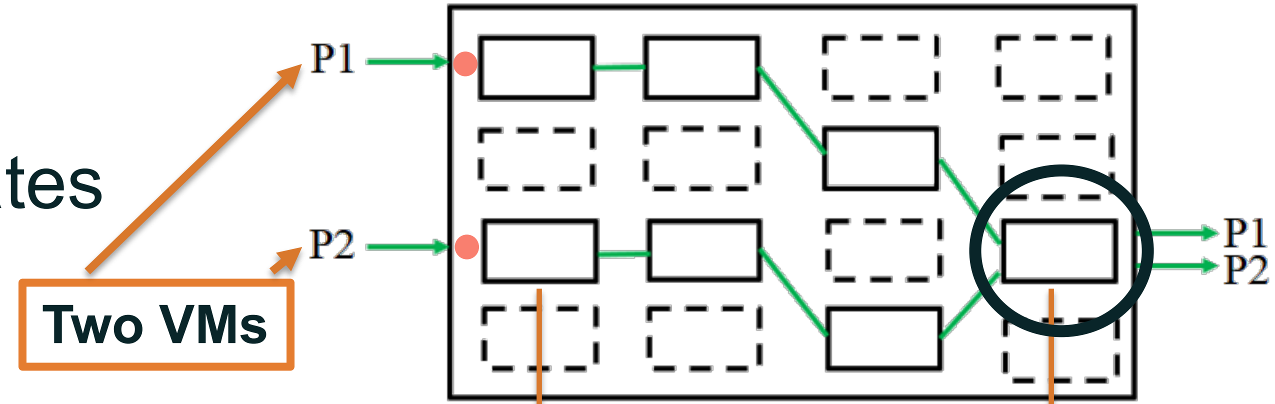
- Connecting several CPU, RAM, and storage boxes using optical switches



[1] G. Zervas, H. Yuan, A. Saljoghei, Q. Chen, and V. Mishra, "Optically disaggregated data centers with minimal remote memory latency: Technologies, architectures, and resource allocation [Invited]," Journal of Optical Communications and Networking, 2018.

Optical Switch Energy Model

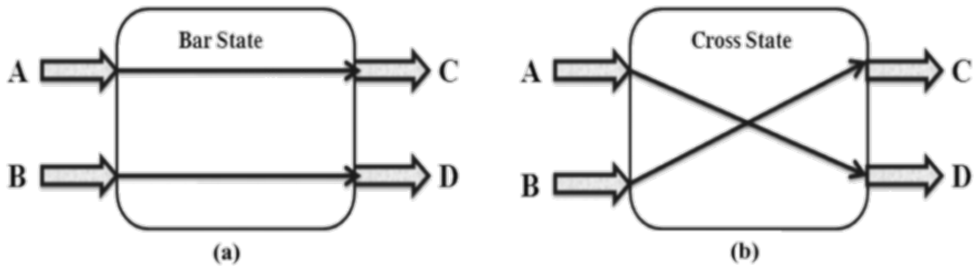
- For low latency and high bandwidth
  - Microring resonator-based switch cells
- Network of cells in bar and cross states
- Energy consumption per VM



Generic view of an optical switch

$$E_{sw} = \left(\frac{n}{2} \times P_{swcell} \times lat_{sw}\right) + (\alpha \times n \times P_{trimcell} \times T)$$

Switching — Maintain state

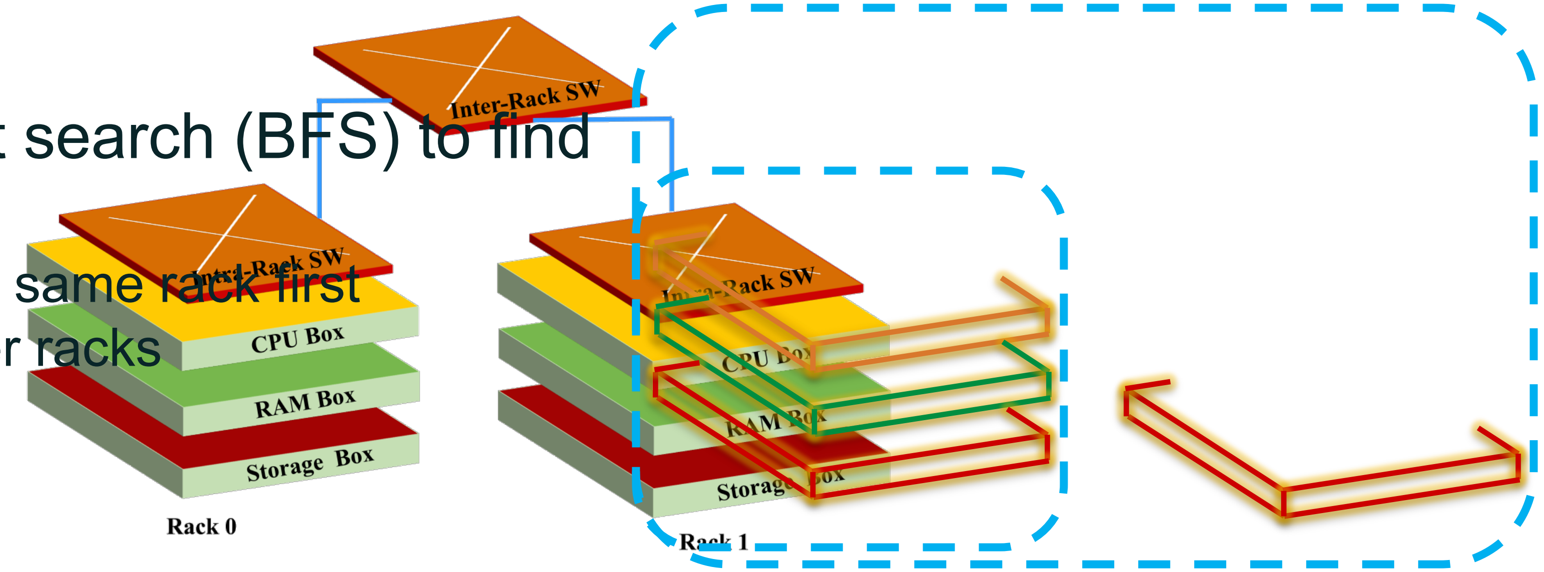


- Here,
  - E_{sw} is the energy per path (or VM)
  - n is the number of cells along a path
  - P_{swcell} is the cell switching power
  - lat_{sw} is the switching latency
  - α accounts for two paths sharing cells
  - $P_{trimcell}$ is the cell trimming power
  - T is the VM lifecycle

Total energy consumption in switch
$= \text{Avg. } E_{sw} \times \text{Number of VMs switched}$

DDC Scheduling Algorithms: NULB [1]

- **N**etwork-**U**naware **L**ocality **B**ased (**NULB**) resource allocation algorithm [1]
- For an incoming VM
  - NULB uses contention ratio (CR)
    - $CR_{CPU} = \frac{\mathrm{CPU_{VM}}}{\mathrm{Total\ Av.\ CPU}}$; $CR_{CPU} > CR_{RAM} > CR_{Storage}$
    - CPU is in highest demand
    - $Rack\ 0\ CPU > CPU_{VM}$
    - Uses breadth-first search (BFS) to find other resources
      - Resources in the same rack first
      - Resource in other racks



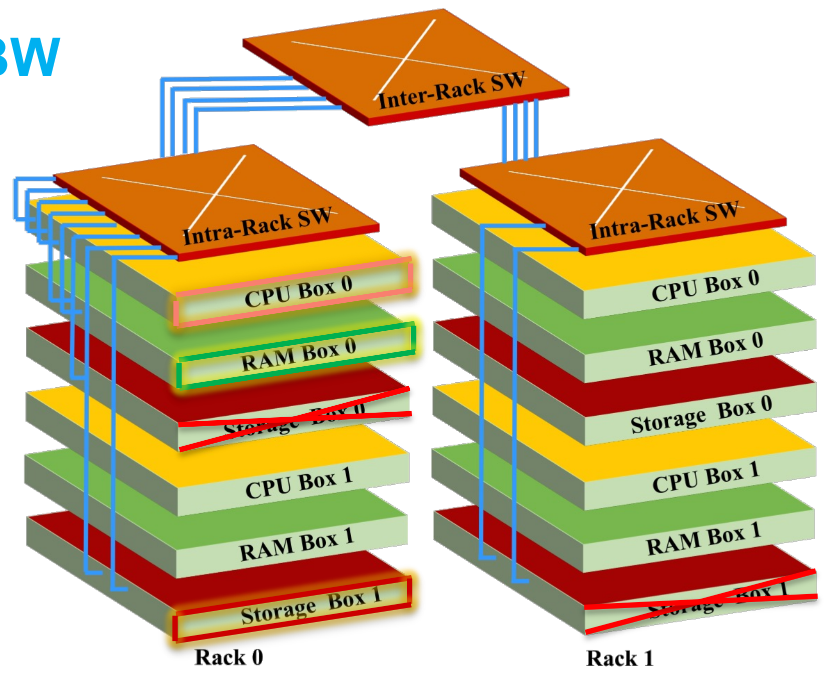
Simplified Case Study Architecture [1]

[1] G. Zervas, H. Yuan, A. Saljoghei, Q. Chen, and V. Mishra, “Optically disaggregated data centers with minimal remote memory latency: Technologies, architectures, and resource allocation [Invited],” Journal of Optical Communications and Networking, 2018.

DDC Scheduling Algorithms: NALB [1]

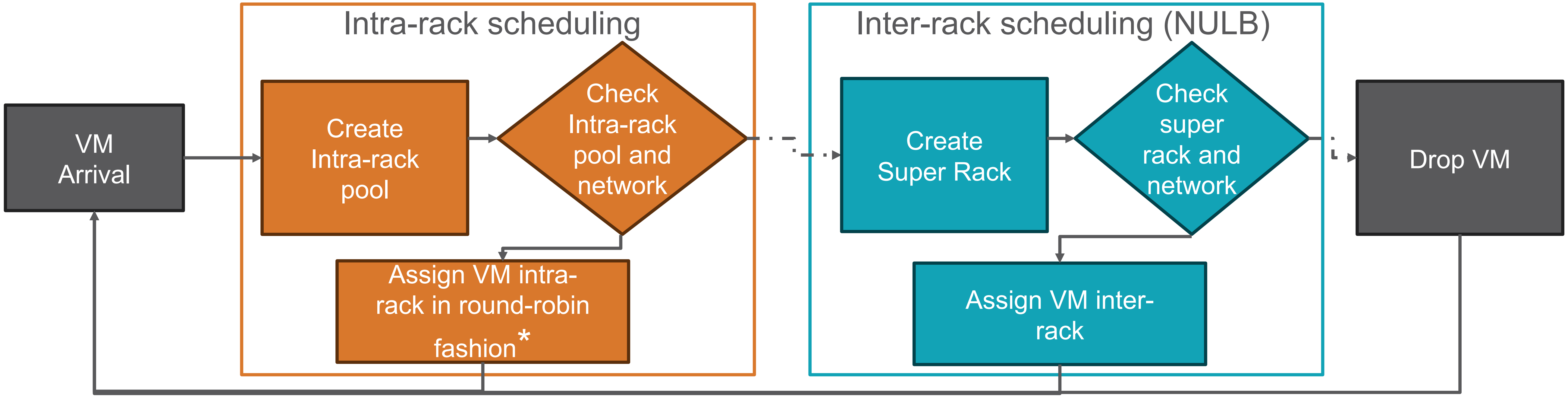
- **N**etwork-**A**ware **L**ocality **B**ased (**NALB**) resource allocation algorithm [1]
- After finding the resource in the highest demand → Modified BFS
  - Neighbors with the most available BW are selected
  - Links with the most available BW are selected

Number of blue links → Av. link BW



[1] G. Zervas, H. Yuan, A. Saljoghei, Q. Chen, and V. Mishra, "Optically disaggregated data centers with minimal remote memory latency: Technologies, architectures, and resource allocation [Invited]," Journal of Optical Communications and Networking, 2018.

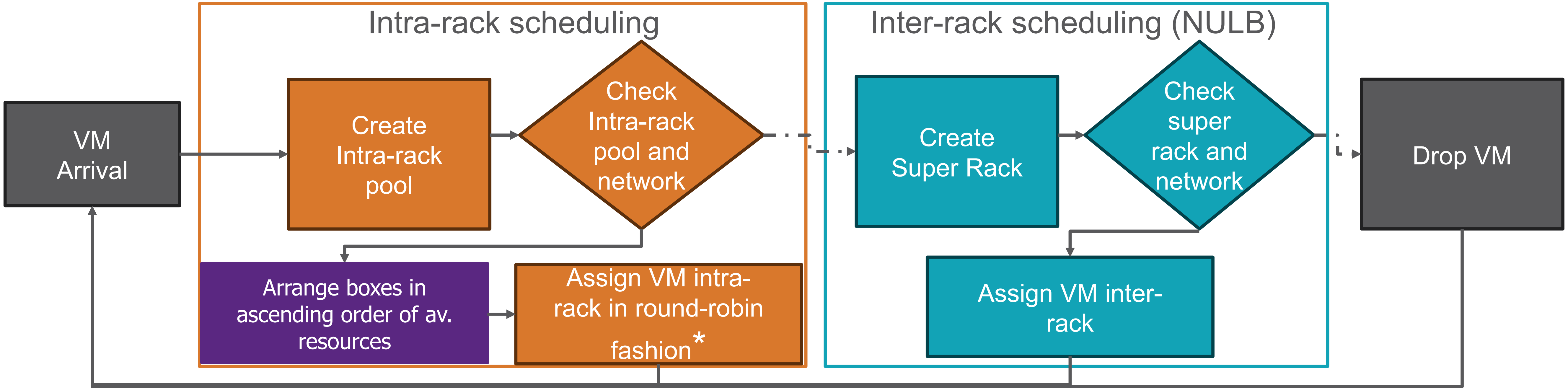
RISA Overview



* Performs Load Balancing [1, 2, 3]

- RISA: Round-Robin Intra-Rack Friendly Scheduling Algorithm
- Main features
  - **Intra-rack pool**: List of racks that can independently schedule a VM
  - **Super rack:** Group of racks that can collectively serve an incoming VM
  - **Load balancing** using Round-robin inspired scheduling

RISA Best-fit (RISA-BF) Overview



Best-fit

- RISA-BF: when the intra-rack pool is not empty
  - Multiple boxes may have sufficient CPU resources
  - RISA-BF will choose the CPU box with the lowest available resources
  - This has been shown to further reduce resource fragmentation

DDC Scheduling Alogoithm Summary

- NULB and NALB implement BFS or Modified BFS
  - This results in high compute resource utilization
  - Highest CR racks often lack other resources
  - More inter-rack VM assignment
    - Sub-optimal network scheduling
    - Increased switch power consumption

- RISA and RISA-BF only perform inter-rack VM assignments to avoid VM drops
  - Fewer inter-rack VM assignments
    - More optimal network scheduling
      - Less switch power consumption
  - Round-Robin → Different sizes of VMs are spread all over
  - Best fitting further reduces resource fragmentation

Experimental Setup

- Synthetic random workload [1]
  - Random sizes of VMs
  - Total of 2500 VMs generated

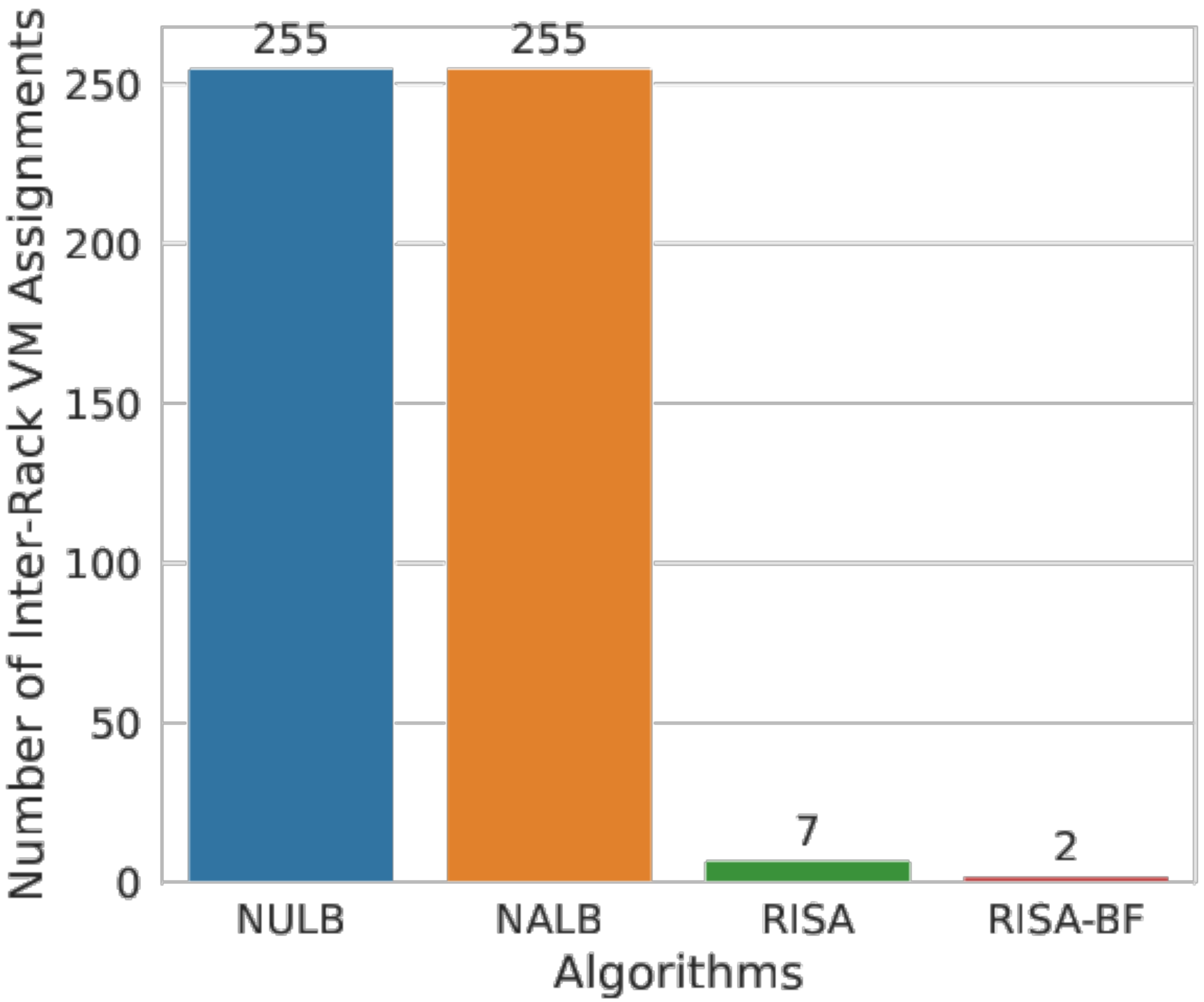
- DDC Configuration
  - Cluster size of 18 racks
  - Rack size of 6 boxes
    - 2 boxes of each kind
  - Three levels of optical switches

[1] G. Zervas, H. Yuan, A. Saljoghei, Q. Chen, and V. Mishra, "Optically disaggregated data centers with minimal remote memory latency: Technologies, architectures, and resource allocation [Invited]," Journal of Optical Communications and Networking, 2018.

Discussion of simulation results

- NULB and NALB use contention ratios to select a rack, which may lack other resource types
  - **More than 10% of VM assignments were inter-rack for synthetic workload**

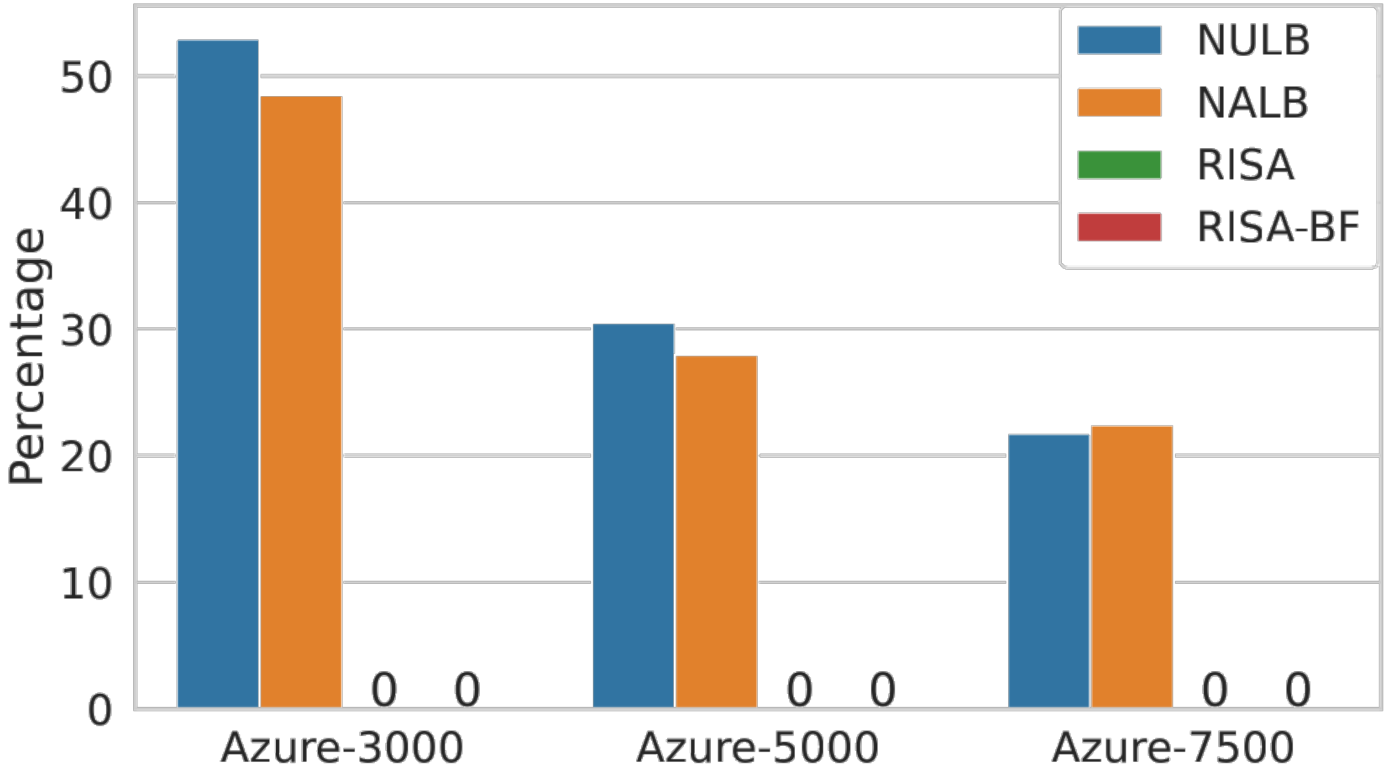
- **RISA** and **RISA-BF** utilized the Intra-rack pool
  - **Less than 1% of VM assignments were inter-rack**
  - **Same compute resource utilization as NULB and NALB**



Inter-rack VM assignment for synthetic workload

Results of using practical workload

- To gauge the performance of RISA in a practical scenario, we used the 2017 Azure data center traces [2]
  - The first 3000 VMs grouped as Azure-3000
  - The first 5000 VMs grouped as Azure-5000
  - The first 7500 VMs grouped as Azure-7500
  - Storage information [1]
- **NULB & NALB**
  - **20% - 50% of VM assignments were inter-rack**
- **RISA and RISA-BF**
  - **NO VM assignments were inter-rack**



Inter-rack VM assignment for practical workload [2]

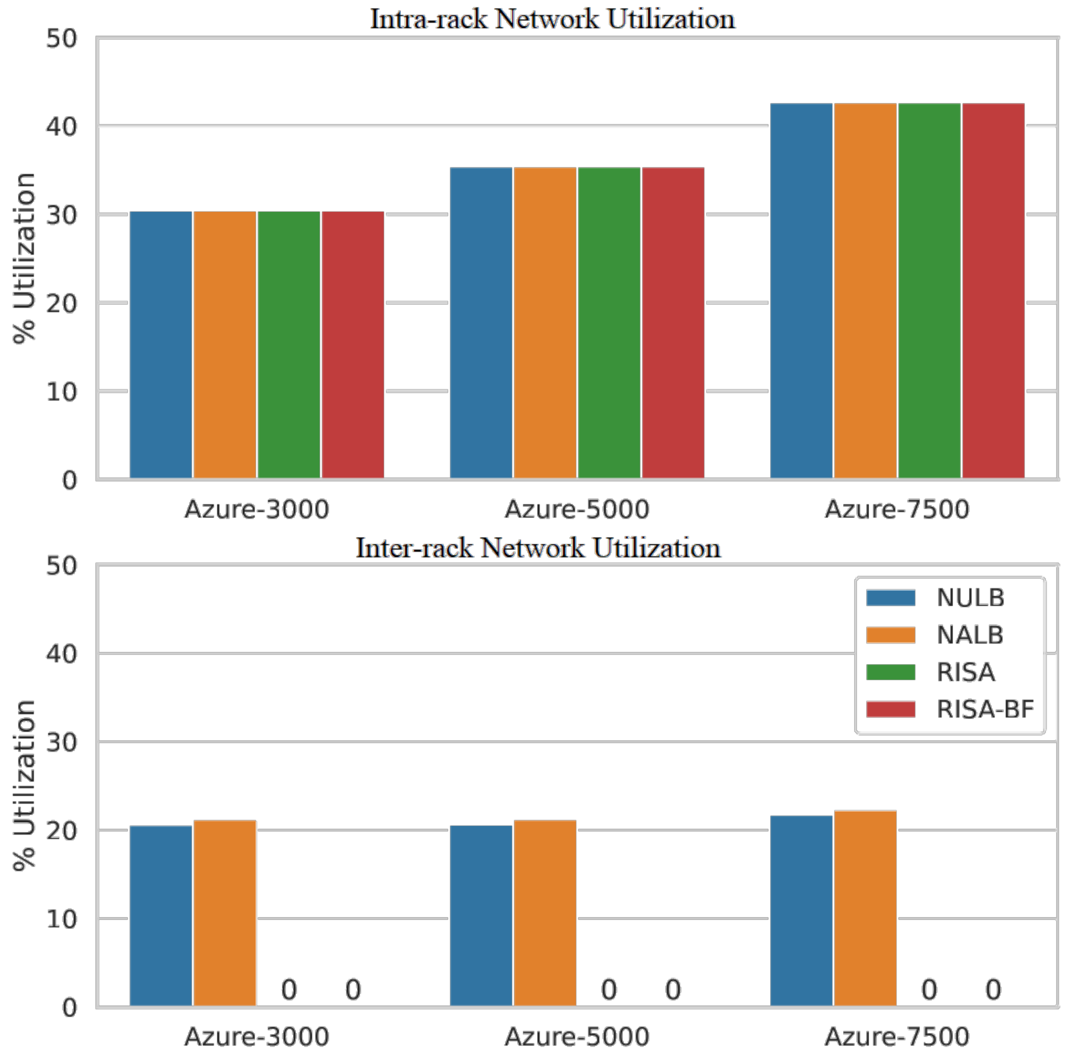
[1] G. Zervas, H. Yuan, A. Saljoghei, Q. Chen, and V. Mishra, “Optically disaggregated data centers with minimal remote memory latency: Technologies, architectures, and resource allocation [Invited],” Journal of Optical Communications and Networking, 2018.

[2] E. Cortez et al., “Resource Central: Understanding and Predicting Workloads for Improved Resource Management in Large Cloud Platforms,” SOSP 2017.

Network utilization

- ➢ Compute resource utilization same for all
  - Intra-rack network used for
    - CPU – RAM communication
    - RAM – storage communication
  - Intra-rack network utilization was also the same

- ➢ RISA and RISA-BF
  - NO inter-rack VM assignment
  - 0% inter-rack network utilization

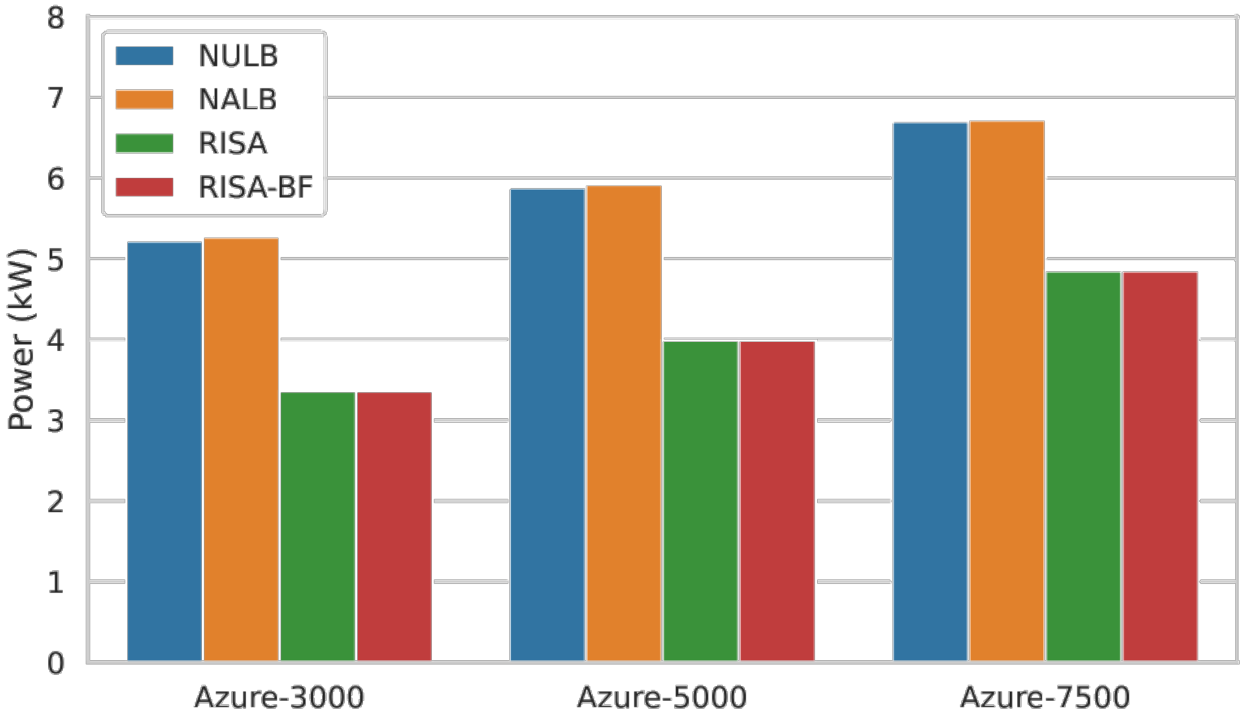


Network utilization for practical workload [2]

[2] E. Cortez et al., “Resource Central: Understanding and Predicting Workloads for Improved Resource Management in Large Cloud Platforms,” SOSP 2017.

Power consumption for optical components

- Transceiver power (22.5 pJ/bit [1]) + total switch power
- Box switch ➡ 64 ports
- Intra-rack switch ➡ 256 ports
- Inter-rack switch ➡ 512 ports
  - For higher connectivity

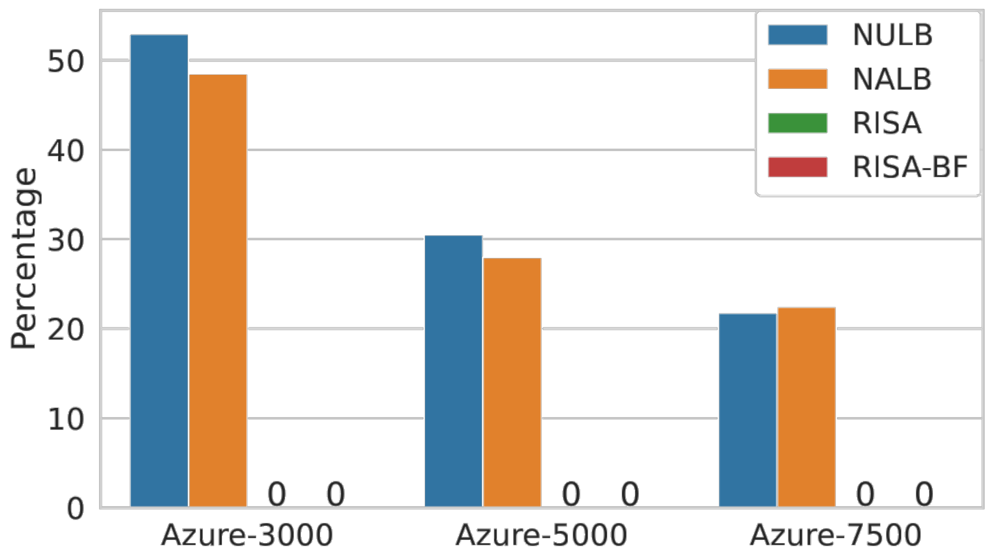


Power consumption for optical components

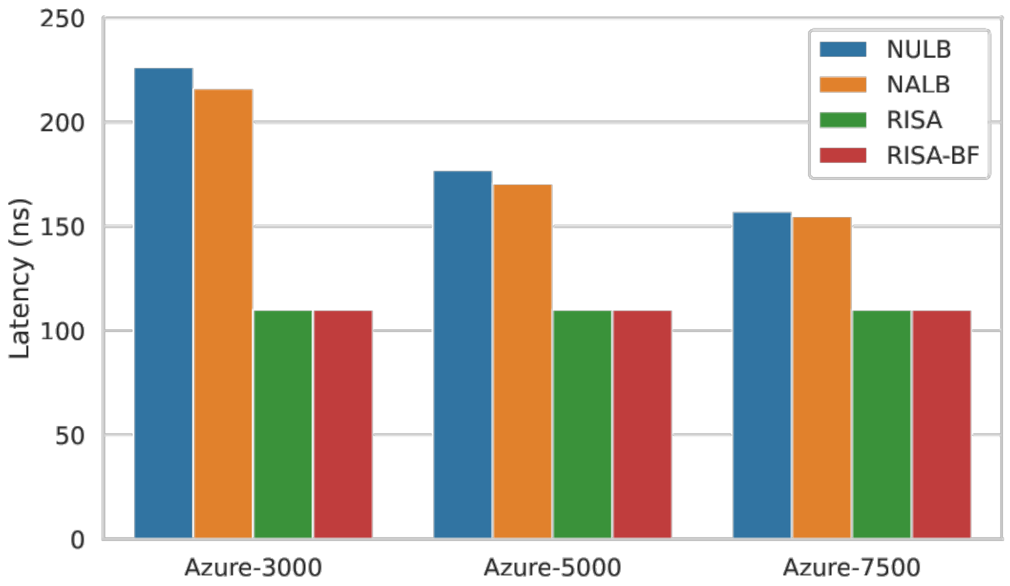
- **RISA** and **RISA-BF**
  - NO Inter-rack network utilization
    - Inter-rack switches consume more power
  - 33% power saving compared to NULB and NALB
  - Power saving will be greater for larger sizes of inter-rack switches

[1] G. Zervas, H. Yuan, A. Saljoghei, Q. Chen, and V. Mishra, "Optically disaggregated data centers with minimal remote memory latency: Technologies, architectures, and resource allocation [Invited]," Journal of Optical Communications and Networking, 2018.

Average CPU-RAM Round-trip Latency



Inter-rack VM assignment for practical workload [2]



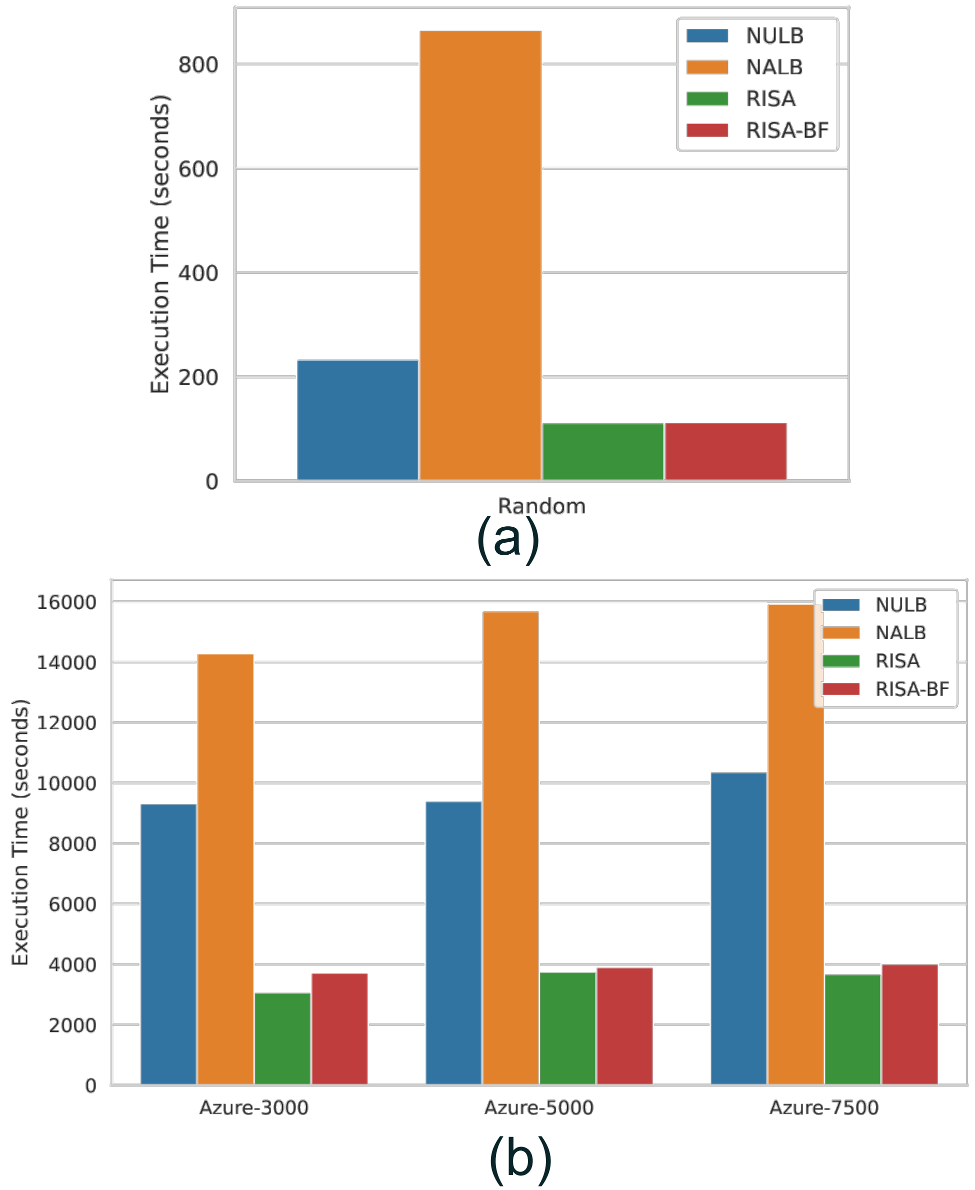
Average CPU-RAM round-trip latency

- NULB average CPU-RAM round-trip latency = 226 ns
- NALB average CPU-RAM round-trip latency = 216 ns
- RISA (or RISA-BF) average CPU-RAM round-trip latency = 110 ns

[2] E. Cortez et al., "Resource Central: Understanding and Predicting Workloads for Improved Resource Management in Large Cloud Platforms," SOSP 2017.

Execution (Scheduling) Time

- NULB, RISA, and RISA-BF
  - Same time complexity
  - Intra-rack pool is empty
    - RISA and RISA-BF use NULB
- In most cases
  - Intra-rack pool was not empty
- For synthetic workload, **RISA and RISA-BF**
  - **2 × speedup compared to NULB**
  - **8 × speedup compared to NALB**
- For practical workload
  - **RISA had 2.81 × speedup for NULB and 4.33 × speedup for NALB**



Execution (scheduling) time for (a) synthetic workload and (b) practical workload

Conclusion

- RISA: Round-Robin Intra-Rack Friendly Scheduling Algorithm for Disaggregated Datacenters
- Prioritizes intra-rack VM assignment
  - More than NULB and NALB
- Performs load balancing to evenly distribute VMs of different sizes
- Best-fit packing for RISA-BF
  - Further improves utilization
- Uses NULB in worst-cases to prevent VM drops

→

- Significant reduction in network usage translates to
  - Up to 33% reduction in power consumption of optical components
  - Up to 50% reduction in CPU-RAM round-trip latency
  - 2.81– 4.33X speedup for practical workload
- Same compute resource utilization as NULB and NALB

Thank you

Rashadul Kabir (rashadul.kabir@colostate.edu)

Colorado State University

Experimental Setup

- Synthetic random workload [1]
  - Random size of VM
    - 1-32 CPU cores, 1-32 GB RAM and 128 GB storage
  - Interarrival rate is based on a Poisson distribution with a mean value of 10 time units
  - VM lifecycle starts at 6300 time units
  - For each set of 100 requests
    - Lifecycle increases by 360 time units
  - 2500 VMs generated

DDC Configuration	
Cluster size	18 racks
Rack size	6 boxes
Box size	8 bricks
Brick size	16 units
CPU unit	4 cores
RAM unit	4 GB
Storage unit	64 GB

[1] G. Zervas, H. Yuan, A. Saljoghei, Q. Chen, and V. Mishra, "Optically disaggregated data centers with minimal remote memory latency: Technologies, architectures, and resource allocation [Invited]," Journal of Optical Communications and Networking, 2018.