Priority-Aware Scheduling for Packet-Switched Optical Networks in Datacenter

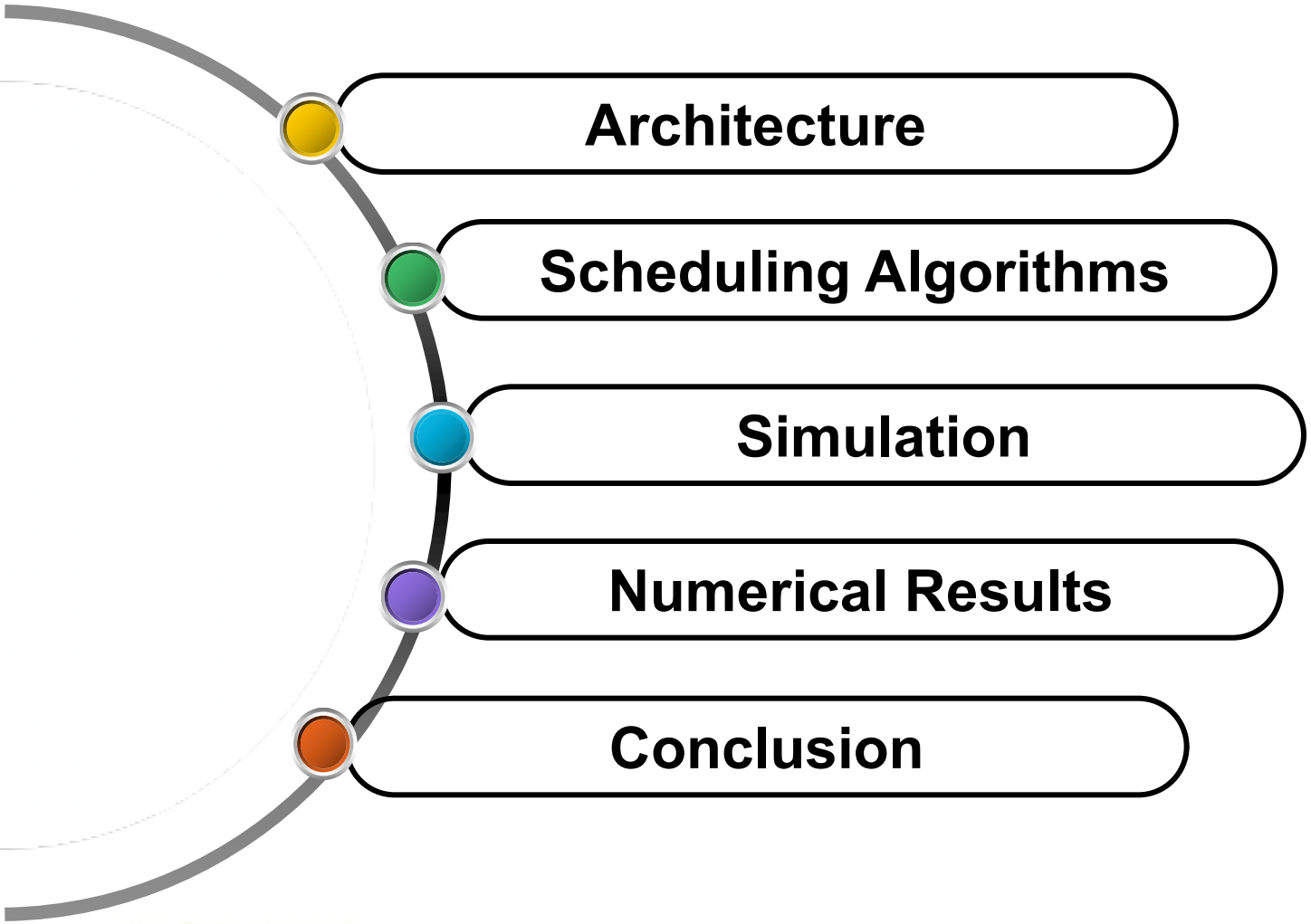
Lin Wang*, Xinbo Wang*, **Massimo Tornatore***†, Kwang joon Kim‡, Sun Me Kim‡,Dae-Ub Kim‡, Kyeong-Eun Han‡, and Biswanath Mukherjee*

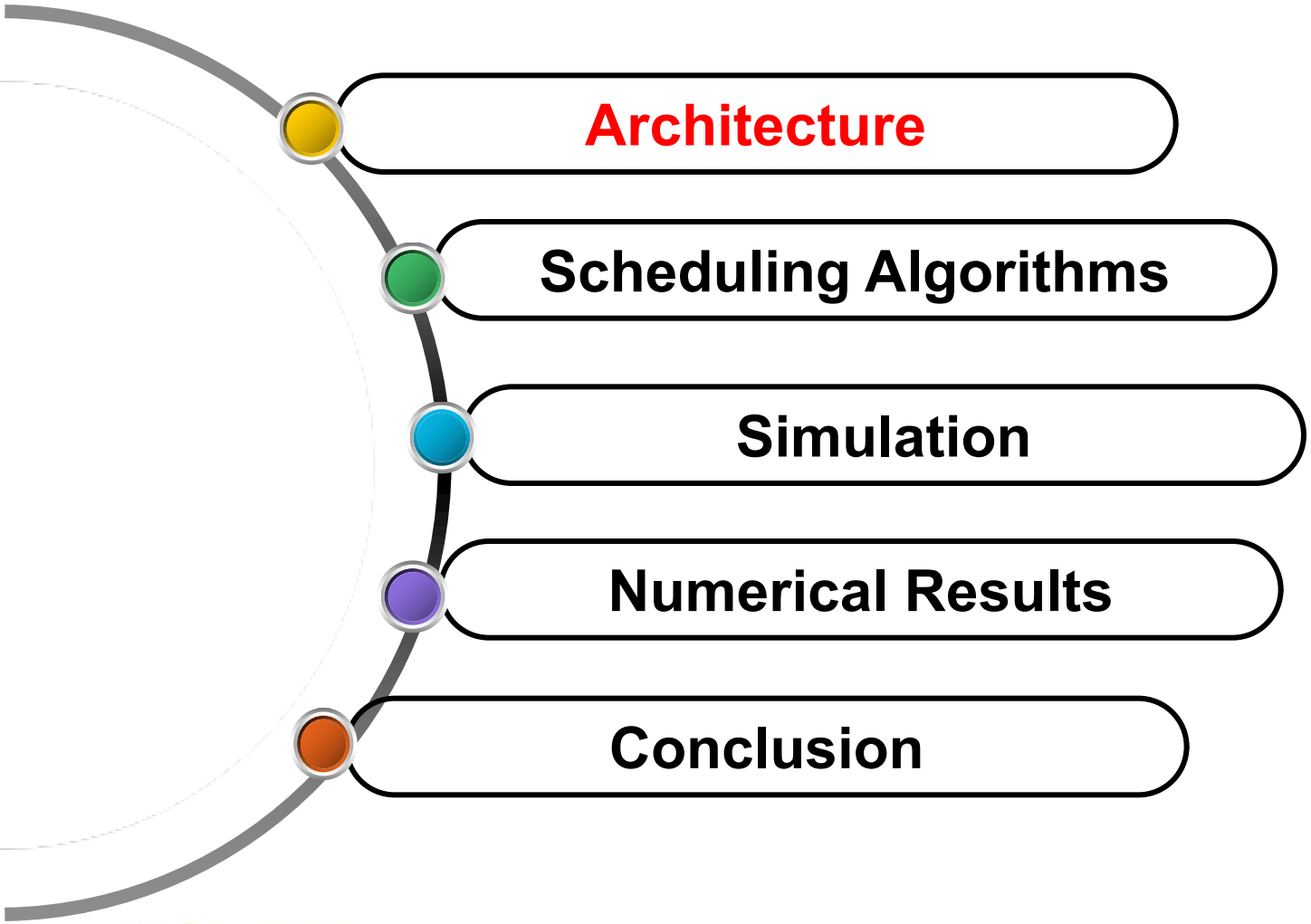
*University of California Davis, USA
†Politecnico di Milano, Italy
‡Electronics and Telecommunications Research Institute, Korea

UCDAVIS

POLITECNICO MILANO

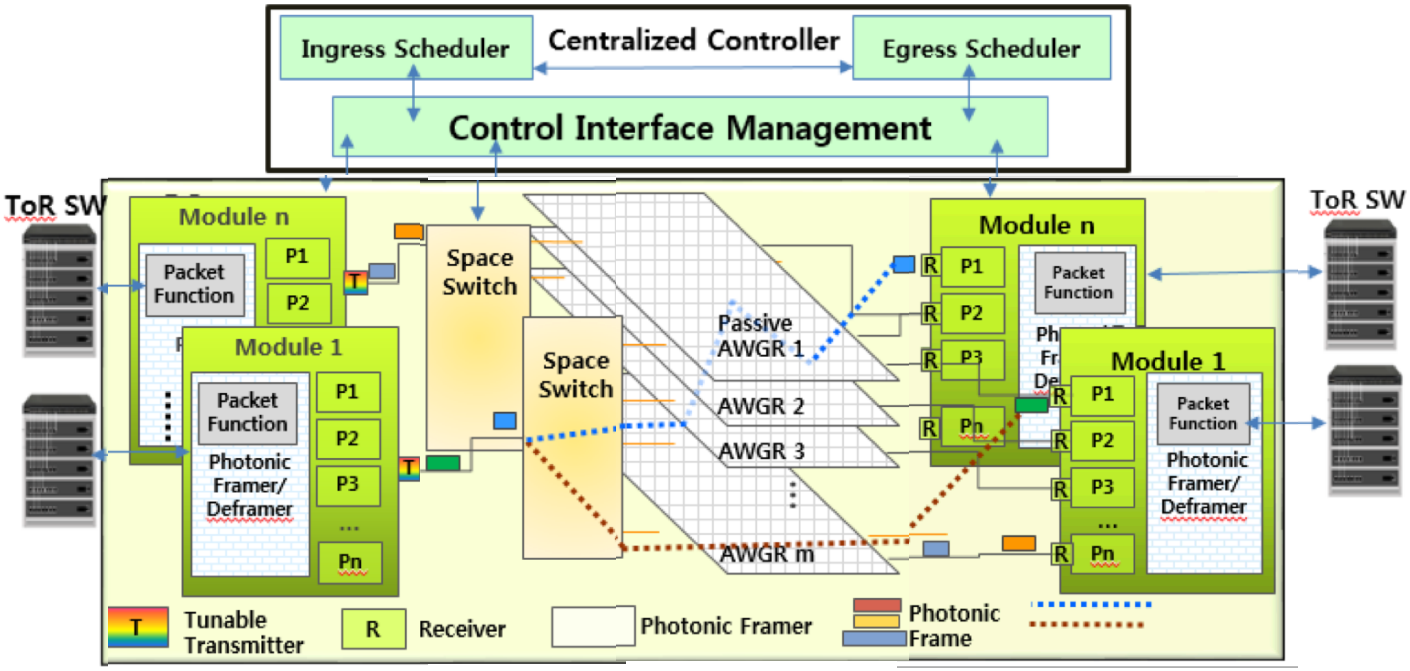
ETRI
한국전자통신연구원
Electronics and Telecommunications Research Institute


Architecture
Scheduling Algorithms
Simulation
Numerical Results
Conclusion


Architecture
Scheduling Algorithms
Simulation
Numerical Results
Conclusion

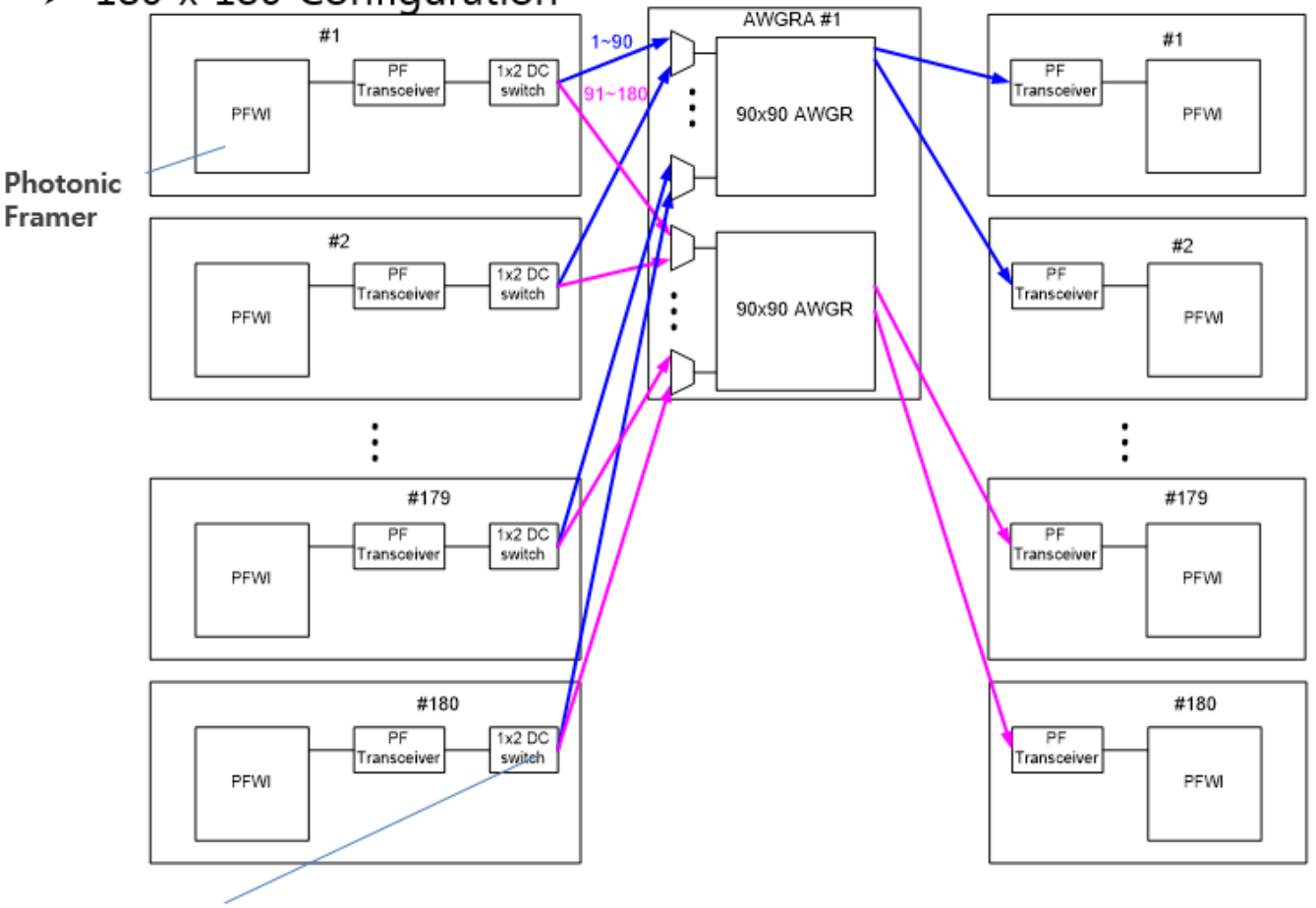
Packet-Switched Optical Network (PSON) Architecture

- Top-of-rack (ToR) switch connected with servers
- Ingress module/egress module with framer/de-framer per switch
- A space switch connects with *m* Arrayed Waveguide Grating Routers (AWGR)
- Scalable core switching consists of AWGRs
- Control interface management

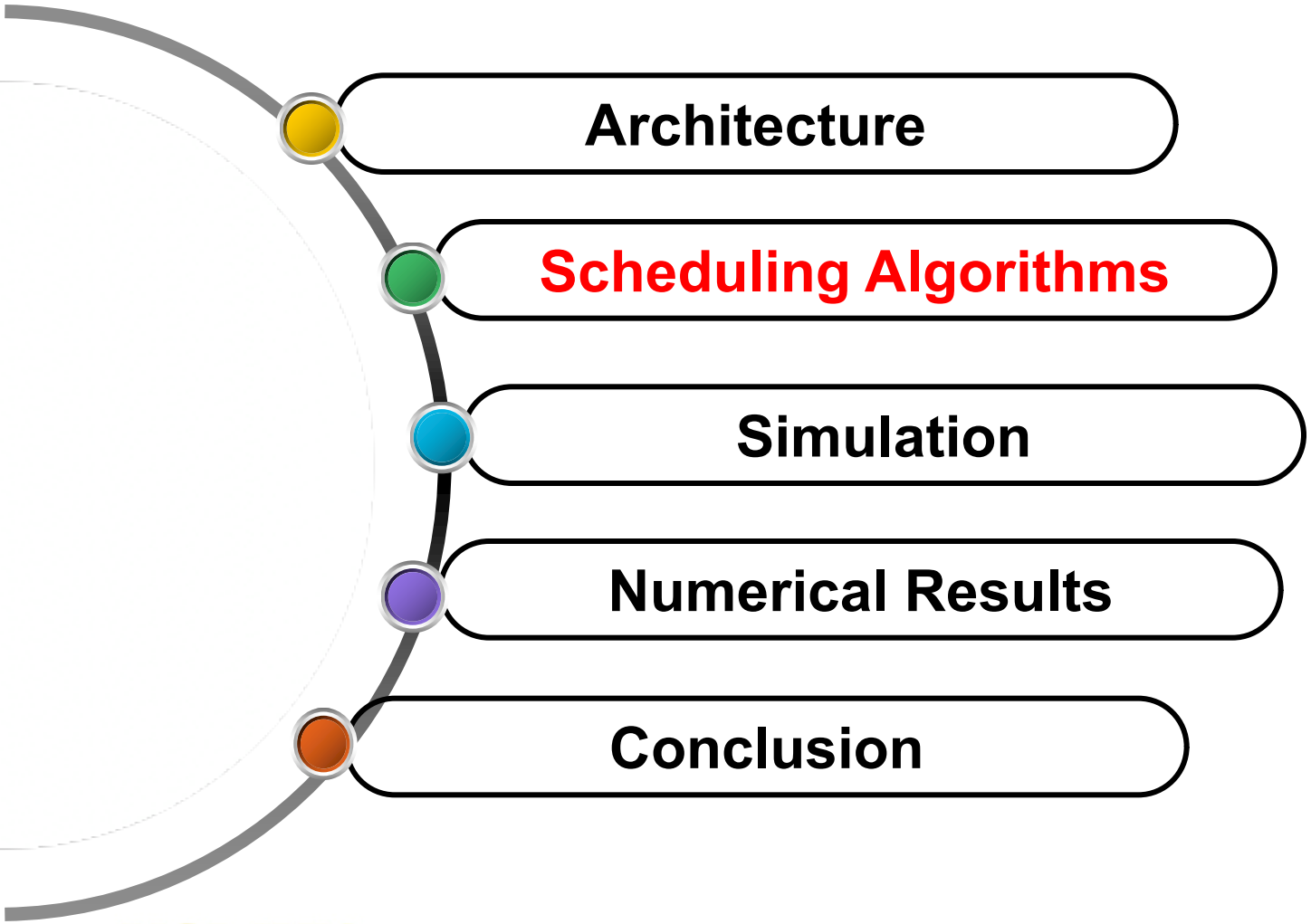


PSON data plane (with optical switch fabric)

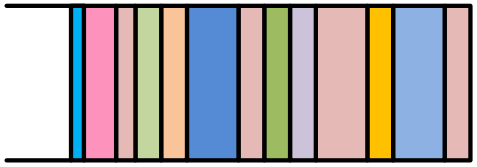
- 180 x 180 Configuration



Space switch: Optical path switch to switch optical signal between module and AWGR at sub micro-second speed.

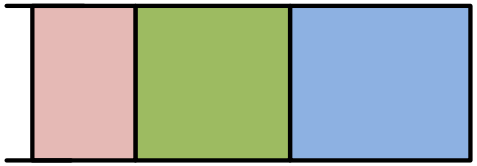

Architecture
Scheduling Algorithms
Simulation
Numerical Results
Conclusion

Mice VS. Elephant Flow



- Small size packet
- Large number
- Short flow
- Short-lived

transactional traffic, web browsing, search queries (≈ 80% traffic)

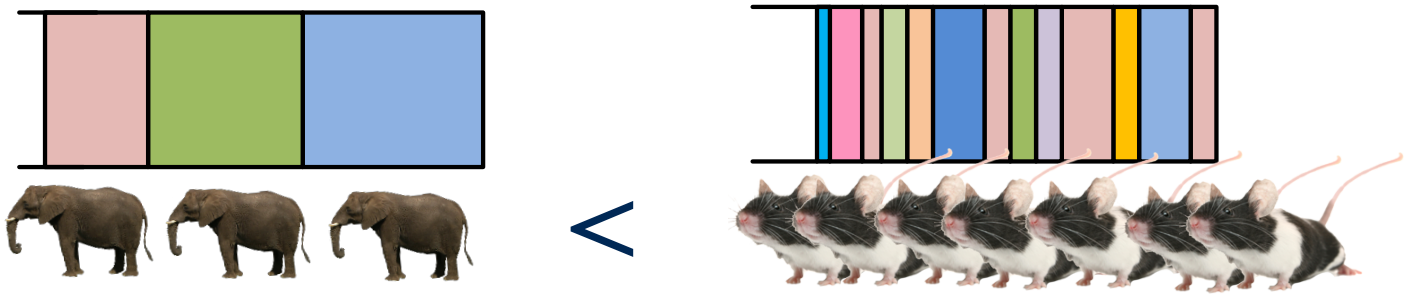


- Large size packet
- Small number
- Large volume flow
- Long-lasting

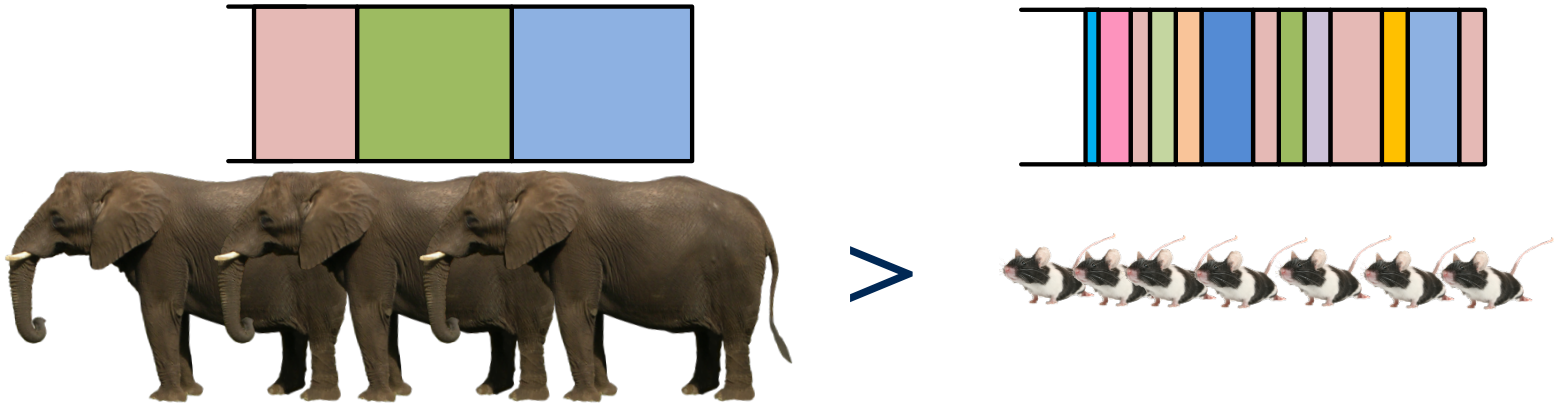
bulk data transfer, data backup, virtual machine migration (≈ 20% traffic)

Mice VS. Elephant Flow

- If schedule first queues with larger **number** of packets, elephant flows will be penalized



- If schedule first queues with larger **total size** of packets, mice flows will be penalized

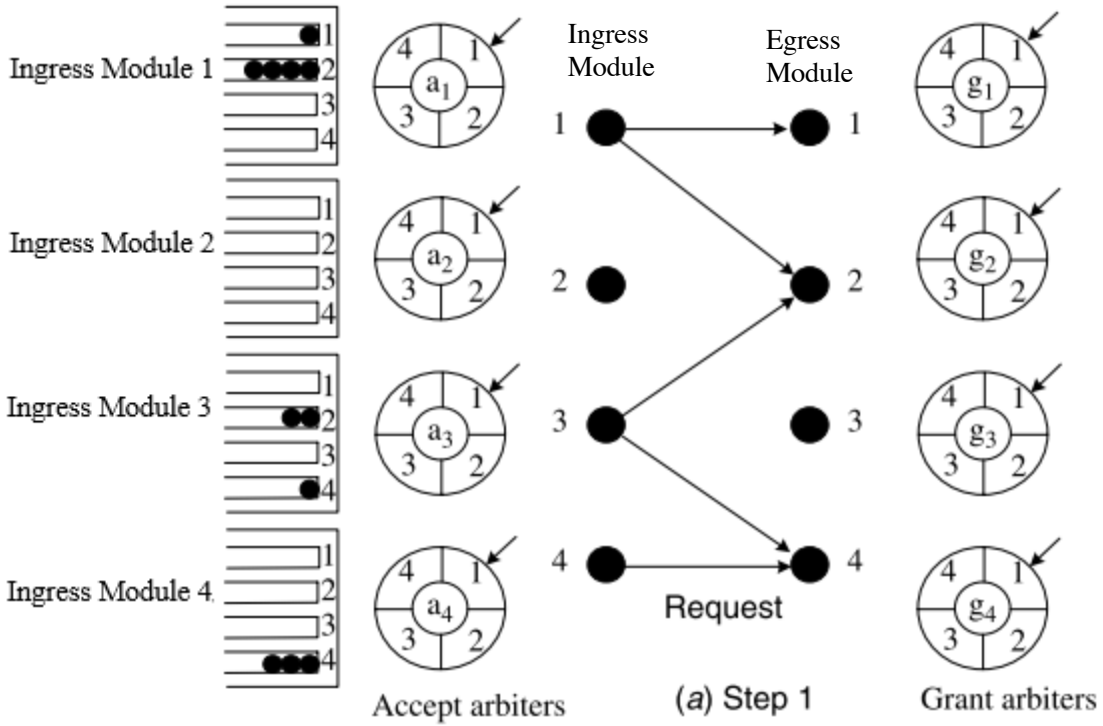


Round Robin Algorithm

- **Iterative Round Robin algorithm**
  - ✓ Step 1: *Request*
  - ✓ Step 2: *Grant*
  - ✓ Step 3: *Accept*

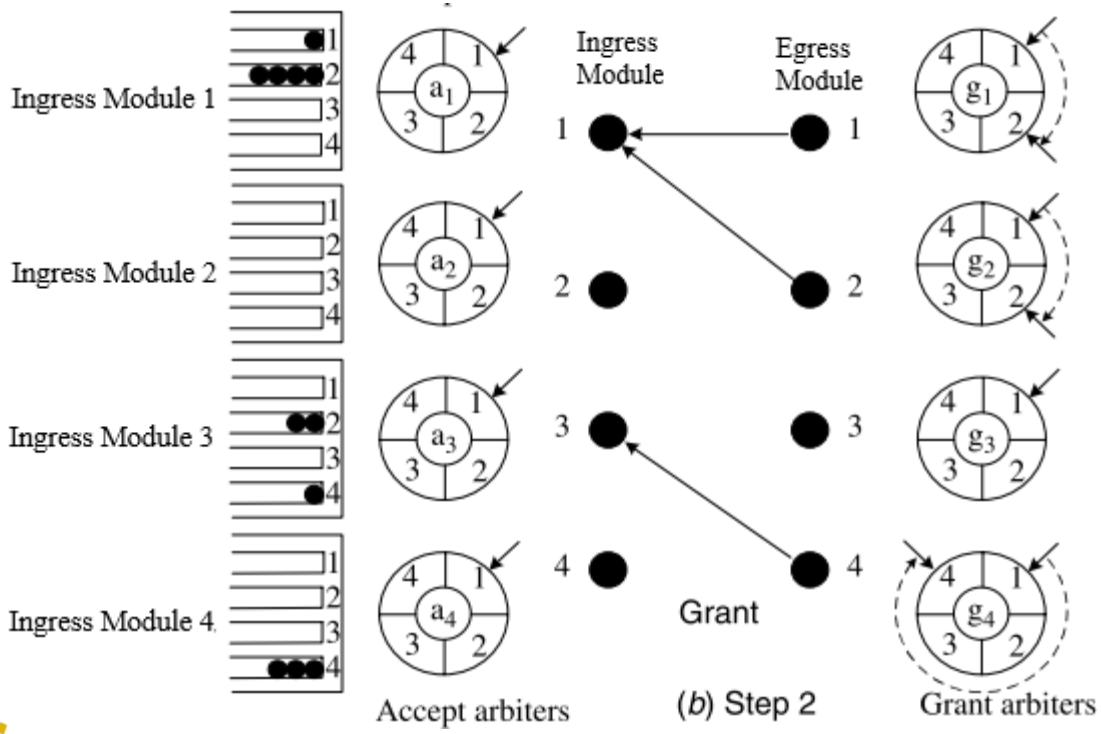
Round Robin Algorithm

- **Step 1. Request**
  - ✓ Each ingress module maintains a queue for each egress module
  - ✓ Send request, if the queue at ingress module is not empty



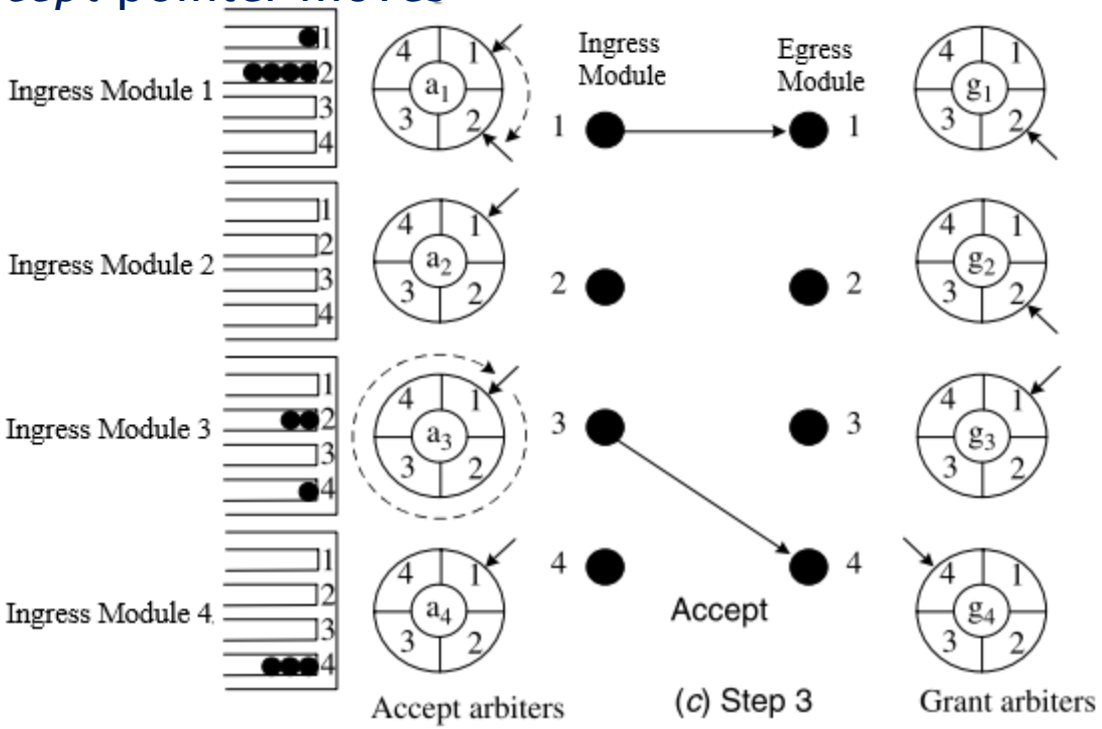
Round Robin Algorithm

- **Step 2. Grant**
  - ✓ Output maintains a *grant* pointer
  - ✓ Output chooses requesting input closest to the grant pointer
  - ✓ *Grant* pointer moves



Round Robin Algorithm

- **Step 3. Accept**
  - ✓ Input maintains a *accept* pointer
  - ✓ Input accepts granting output closest to the accept pointer
  - ✓ *Accept* pointer moves



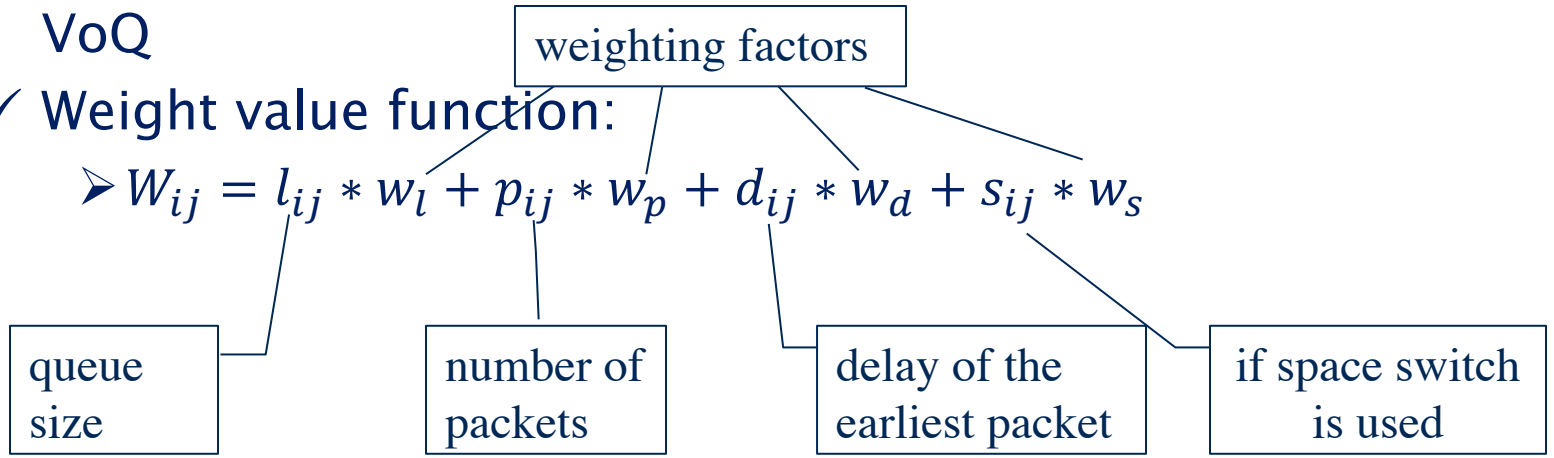
Round Robin Algorithm

- Pro's
  - Simple
  - Fast

- Con's
  - does not consider traffic characteristics
  - results in high delays and unfairness

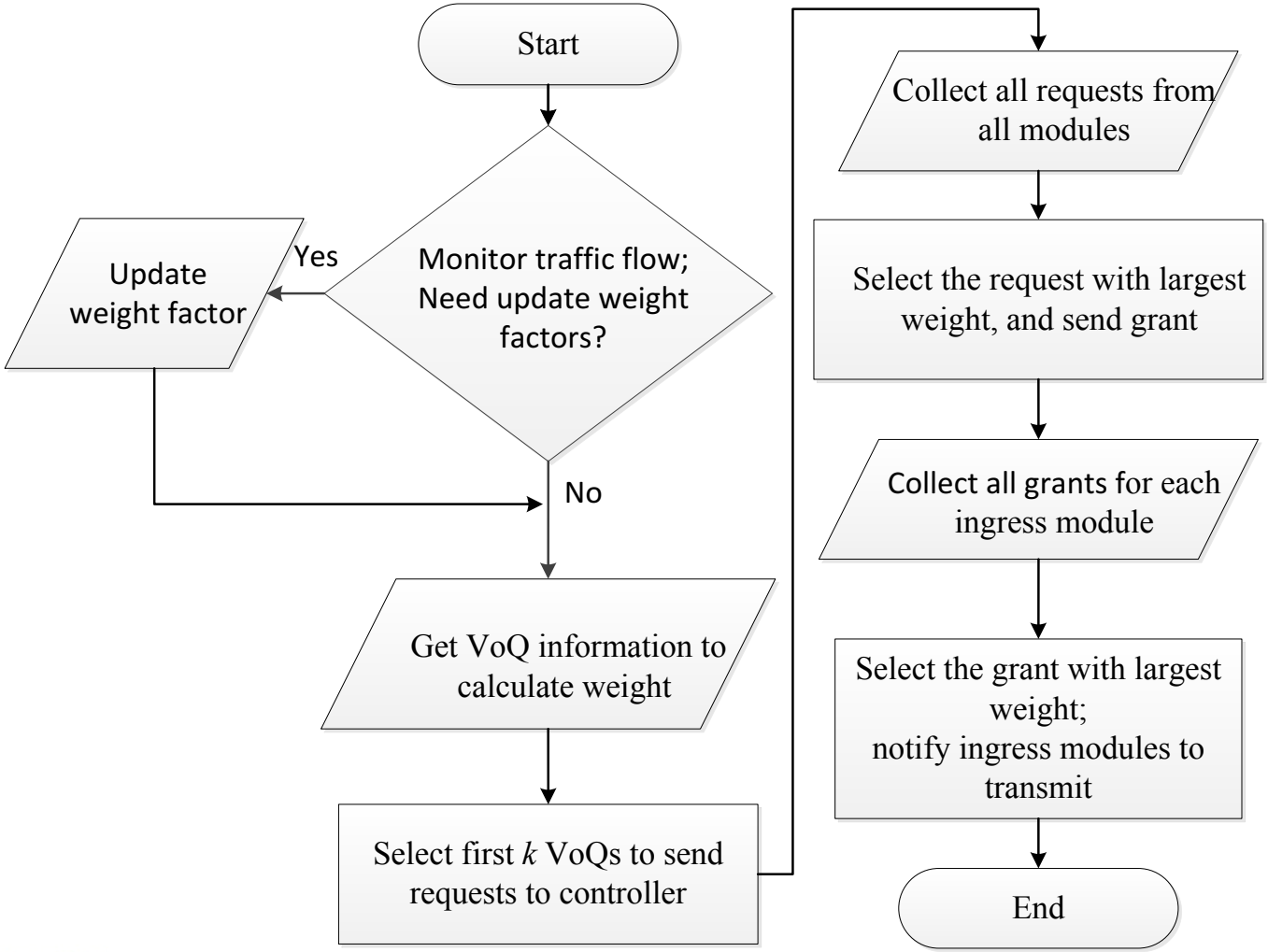
A Priority-Aware Scheduling Algorithm for PSON

- ✓ Uses status information to calculate weight values for each VoQ
- ✓ Weight value function:
  - $W_{ij} = l_{ij} * w_l + p_{ij} * w_p + d_{ij} * w_d + s_{ij} * w_s$

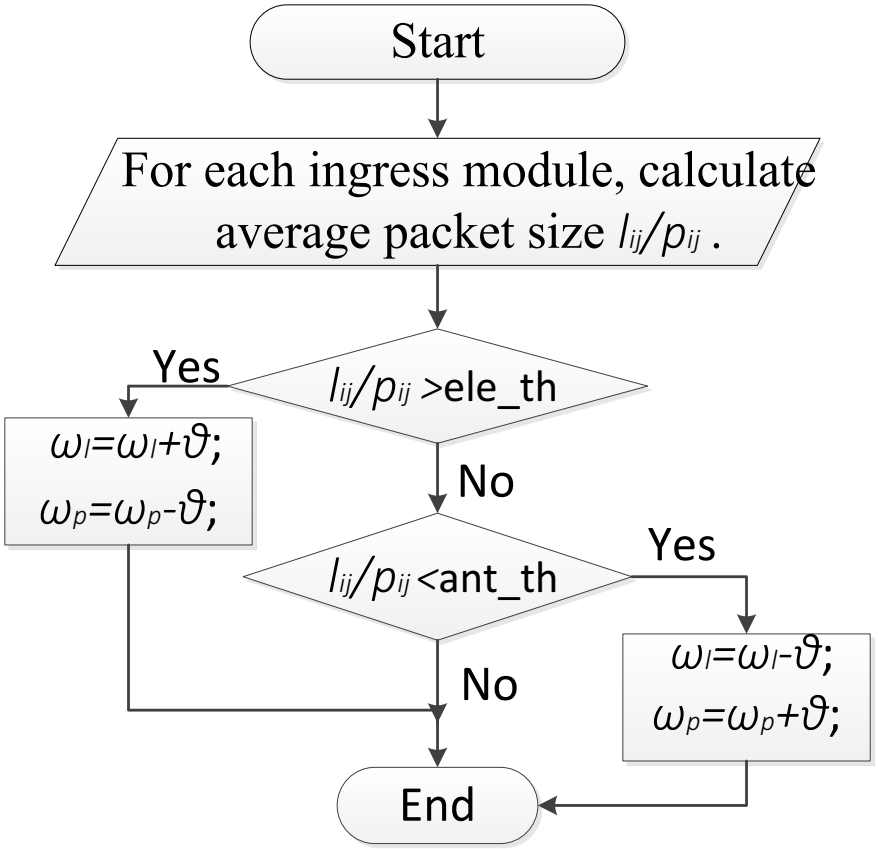


- ✓ Send only first k VoQs with highest weight

Scheduling Algorithm for PSON



Scheduling Algorithm for PSON

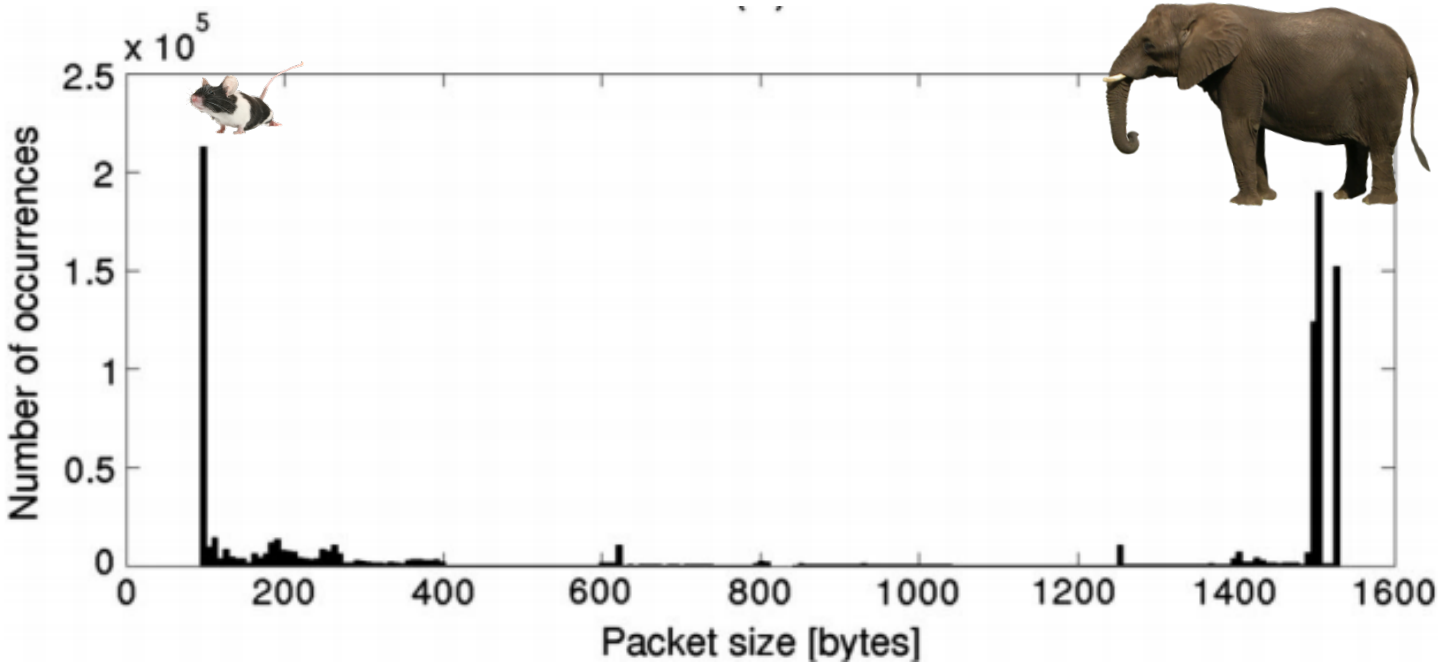


Topology for Simulation

- **PSON:** *n=80* ToR switches, *40×40* AWGRs, and *1×2* space switches.
- **Each ToR receives input traffic generated by 36 servers.**
- **Classify 36 servers of each ToR into three groups:**
  - **1st group: generates packets which contain 80% elephant flow and 20% mice flow;**
  - **2nd group generates packets with 20% elephant flow and 80% mice flow;**
  - **3rd group generates packets with 50% probability.**

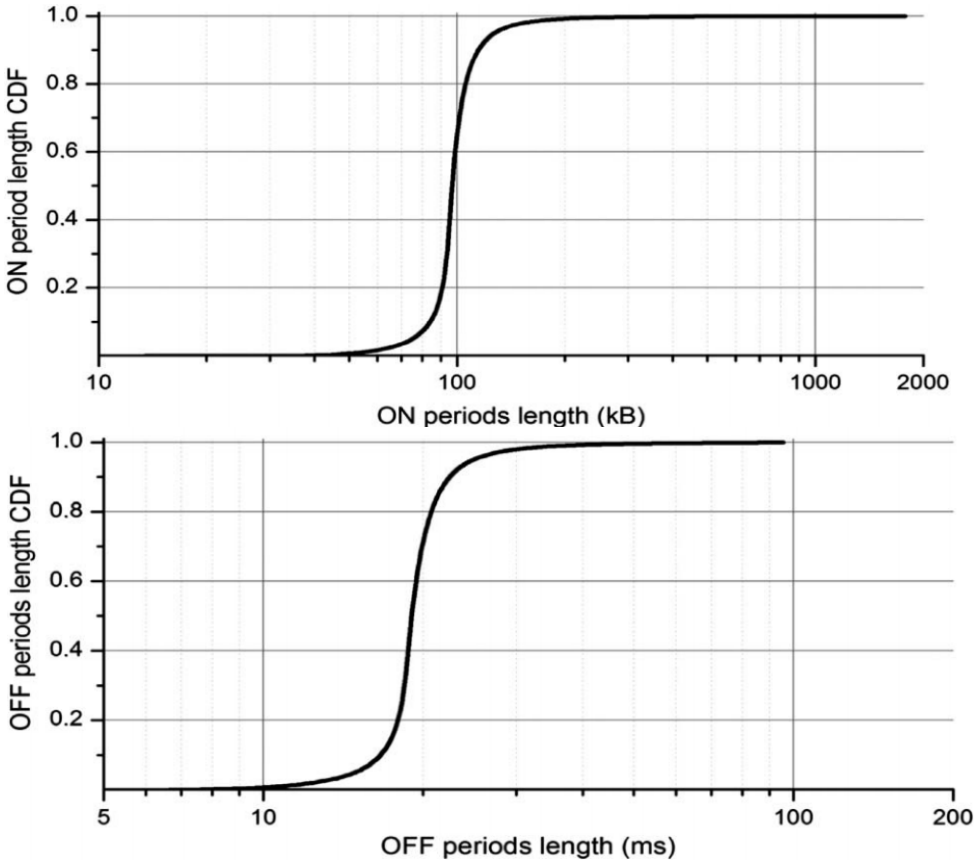
Traffic Generation (1)

- Packet length follows a bimodal distribution around 40 bytes and 1500 bytes



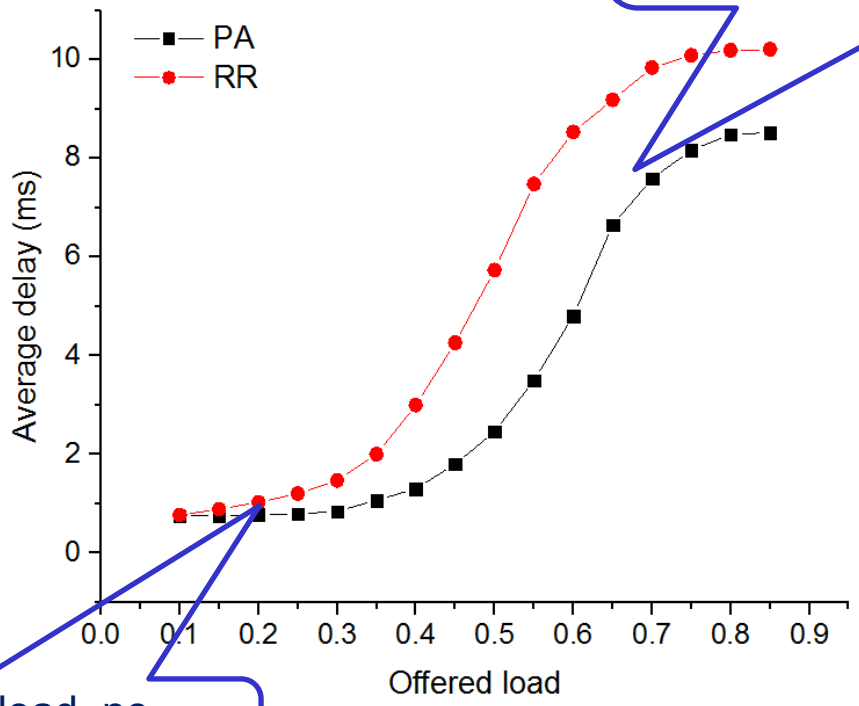
Traffic Generation (2)

- Packet arrival times are modeled matching ON/OFF periods
- ON/OFF periods follows Pareto distribution



Results: Average Delay

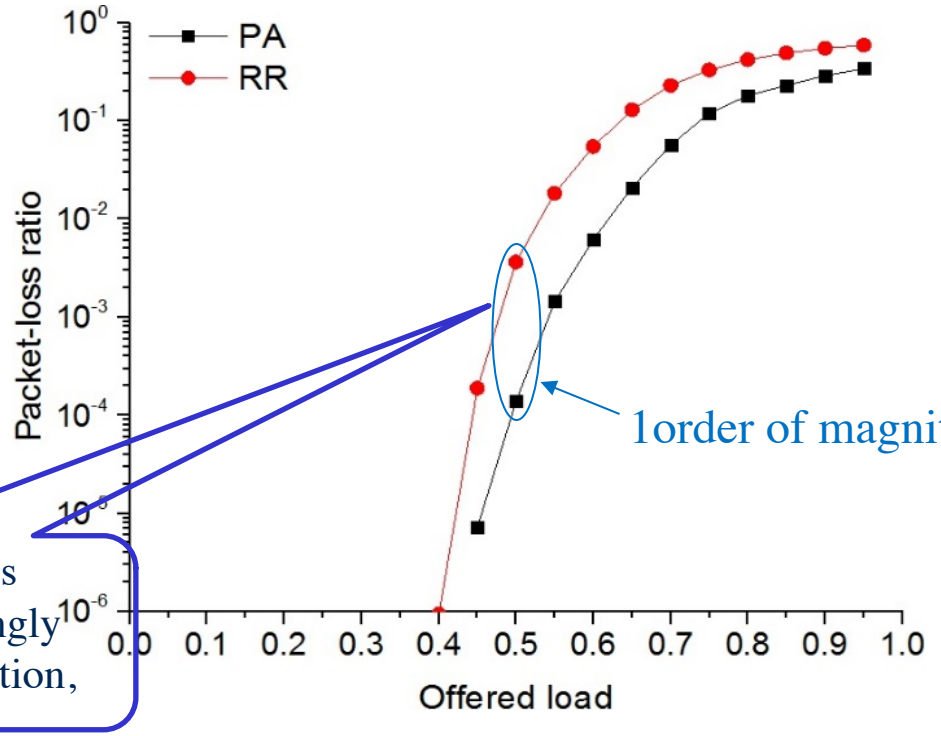
Take-Away 2: Significant delay decrement (more than 50% for intermediate loads)



Take-Away 1: at low load, no difference

Average delay of PA and RR algorithms

Results: Packet Loss Ratio

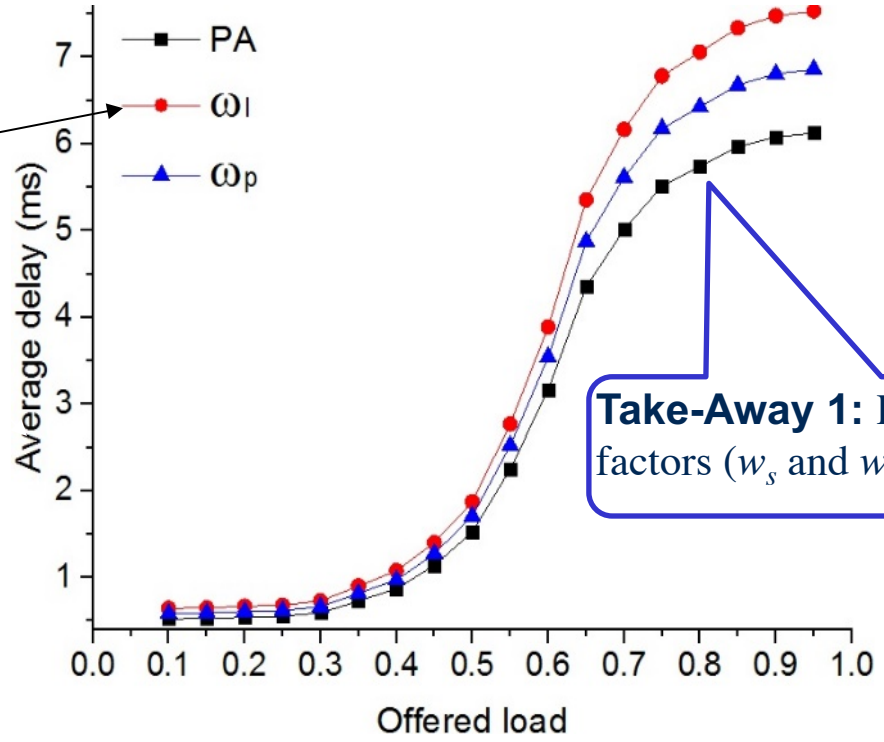


Take-Away 2: PA achieves smaller PLR, and thus strongly improves bandwidth utilization,

Packet-loss ratio of PA and RR algorithms

Results: Average Delay (comparison of ω)

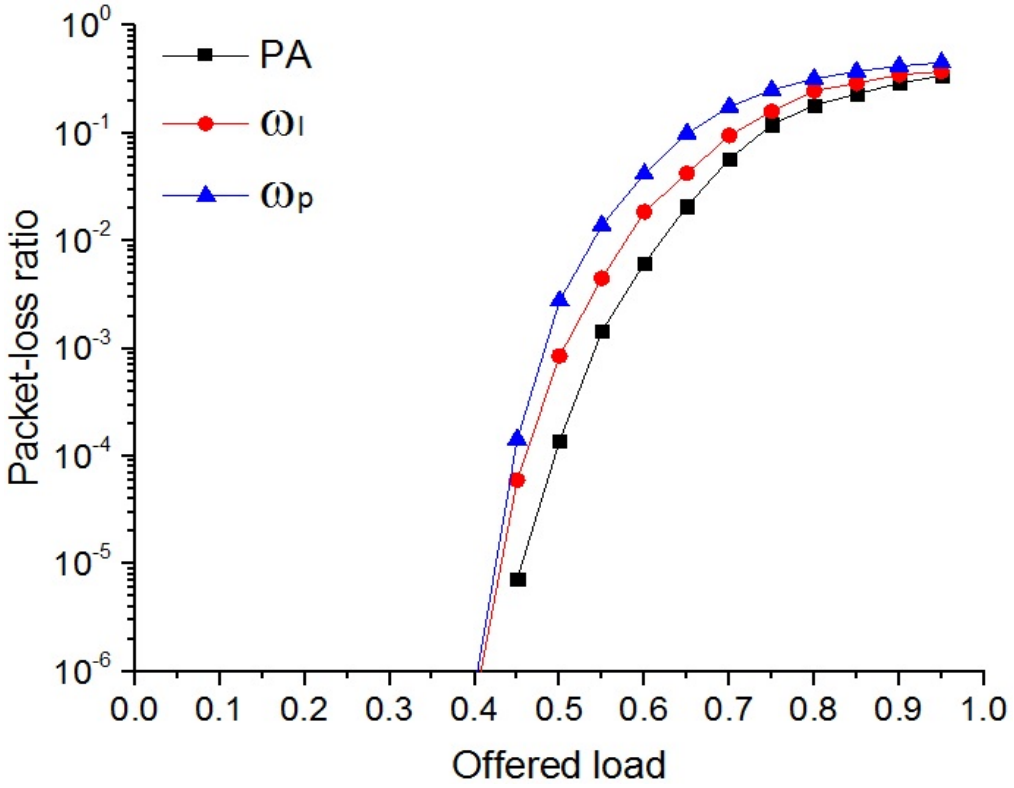
All weight factors are set to zero except w_l, (similarly for the curve of "w_p")



Take-Away 1: PA considering both factors (w_s and w_d) performs the best

Influence of weight factors on average delay of PA algorithms

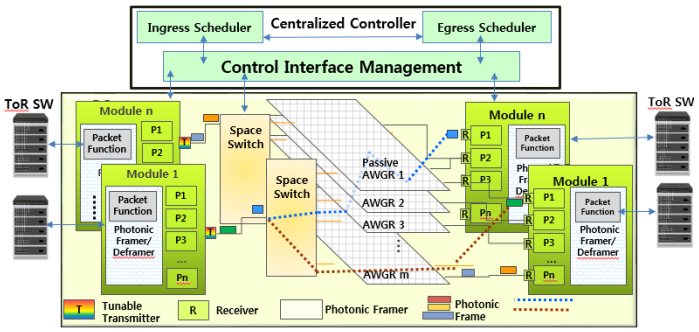
Results: Packet Loss Ratio (comparison of ω)

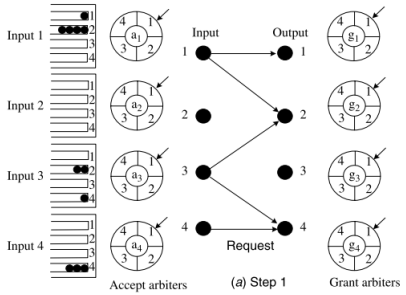


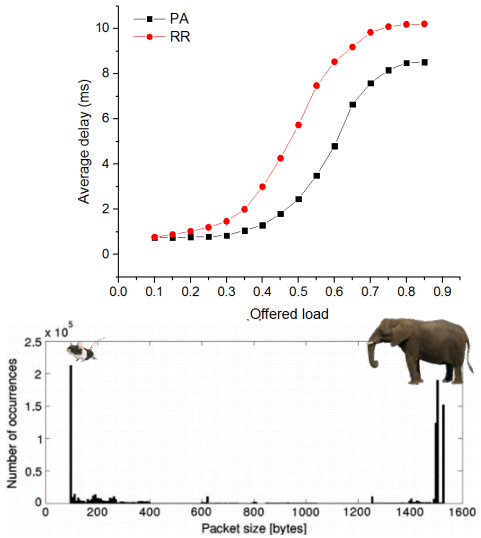
Influence of weight factors on packet-loss ratio of PA algorithms

Conclusion

- We studied scheduling for PSON architecture
- Proposed a Priority-aware (PA) scheduling algorithm
  - Extension of Round Robin to account for «mice vs elephant» flows
- Performed performance evaluation of PA algorithm
- Significant savings: up to one order of magnitude on blocking and up to 50% on delay







UCDAVIS
UNIVERSITY OF CALIFORNIA

amlwang@ucdavis.edu

Scheduling Algorithm for PSON