System-Level Exploration of In-Package Wireless Communication for Multi-Chiplet Platforms

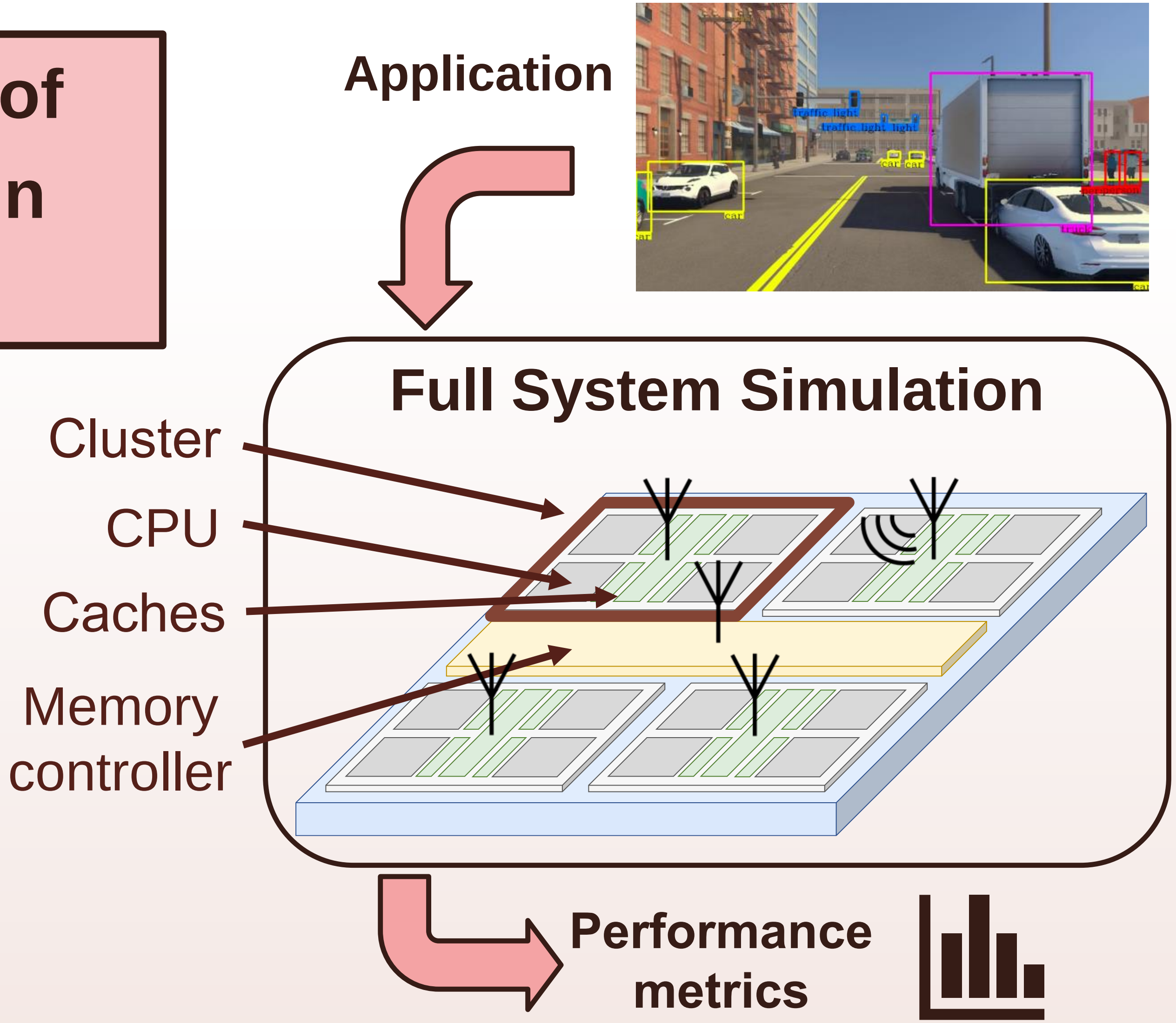
Rafael Medina[1], Joshua Klein[1], Giovanni Ansaloni[1], Marina Zapater[2],
Sergi Abadal[3], Eduard Alarcón[3], David Atienza[1]

[1]Embedded Systems Laboratory (ESL), EPFL, Switzerland,
[2]REDS Institute, HES-SO, Switzerland,
[3]NaNoNetworking Center in Catalonia (N3Cat), UPC, Spain

Motivation

Full-system exploration of wireless interconnects in multi-core systems

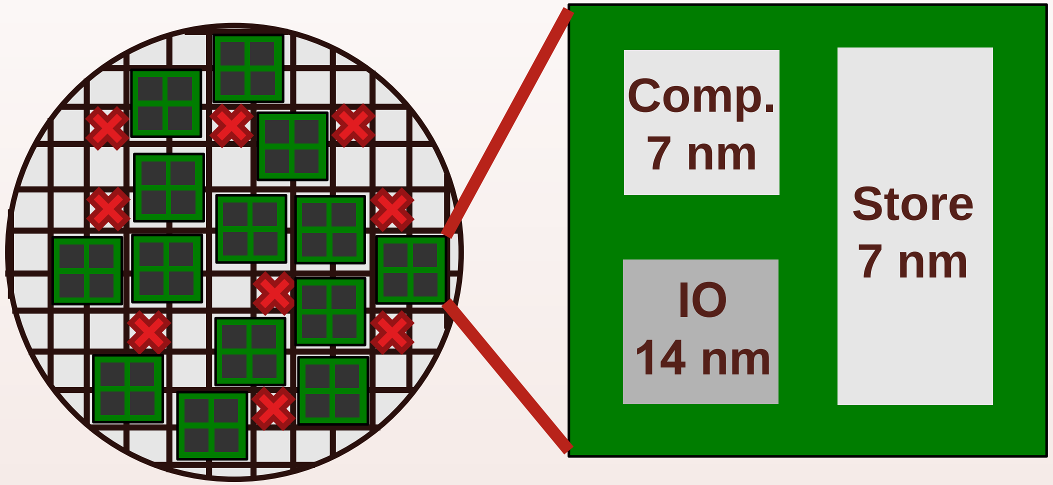
- System architecture
- Wireless in-package communication
- Applications



Multi-Chiplet Platforms

- Large chips
  - Low yield
  - High integration effort

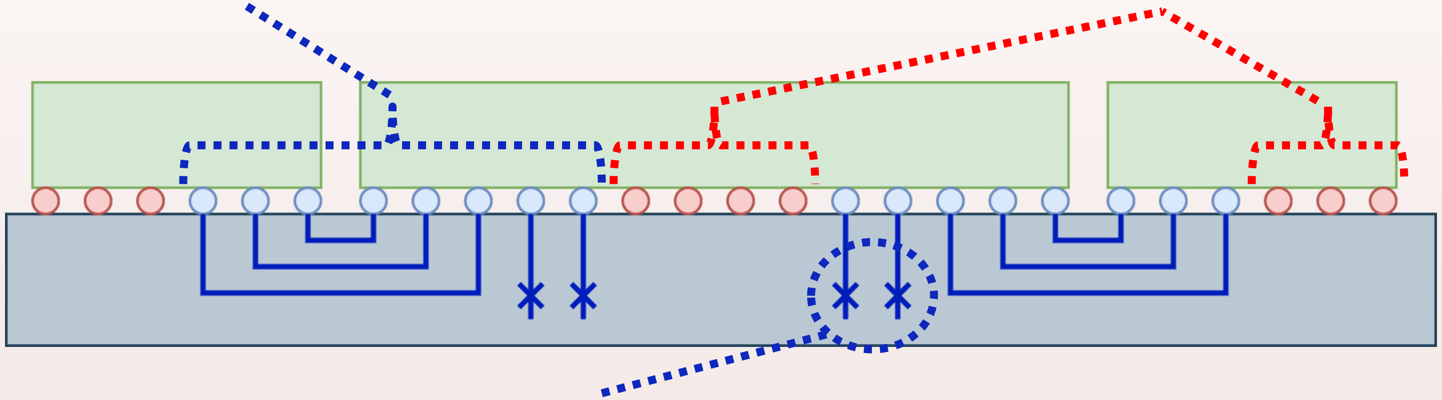
- Solution: **Chiplet + interposer [1]**
  - Several small dies in the same package
  - Increase yield
  - Integration of heterogeneous components



[1] G. Loh et al., DATE, 2021

Chiplet Challenges

- Limited number of chiplet microbumps → { Connectivity, Power delivery }
- Integration of heterogeneous SoCs → Inter-chiplet communication



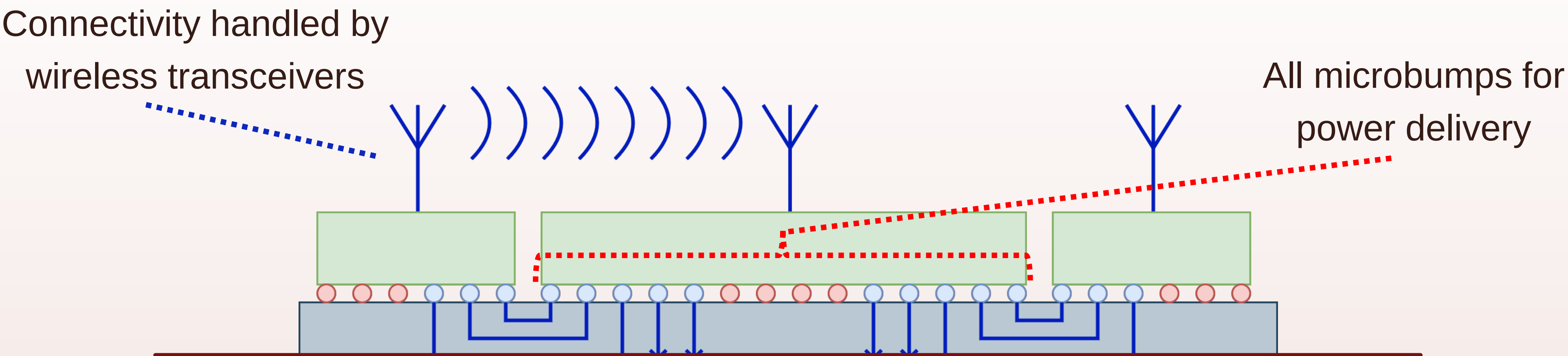
On-substrate inter-chiplet connection

Power supply microbumps

Incompatible interface between chiplets

Solving the Chiplet Interconnect Problem

- Short distance wireless communication [2]
  - Up to 120 Gbps demonstrated
- Addresses chiplet challenges



Connectivity handled by wireless transceivers

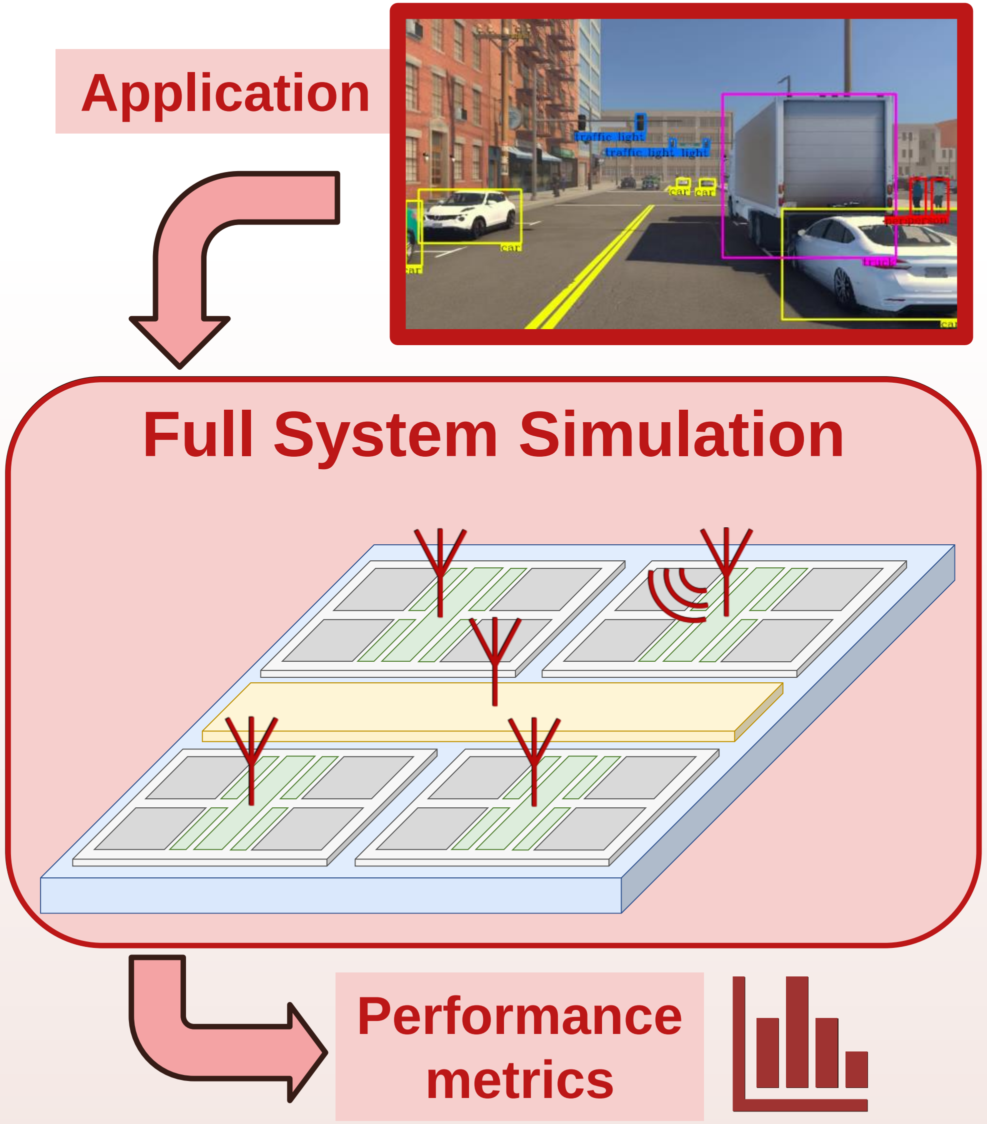
All microbumps for power delivery

We need to assess wireless interconnects

[2] K. Tokgoz et al., ISSCC, 2018

Contributions

- Provide full system framework supporting in-package wireless
- Explore system-level performance of inter-chipet wireless connectivity
- Compare wireless and wired chiplet interconnects
  - Executing complex applications



Application

Full System Simulation

Performance metrics

Wireless in Full System Simulation

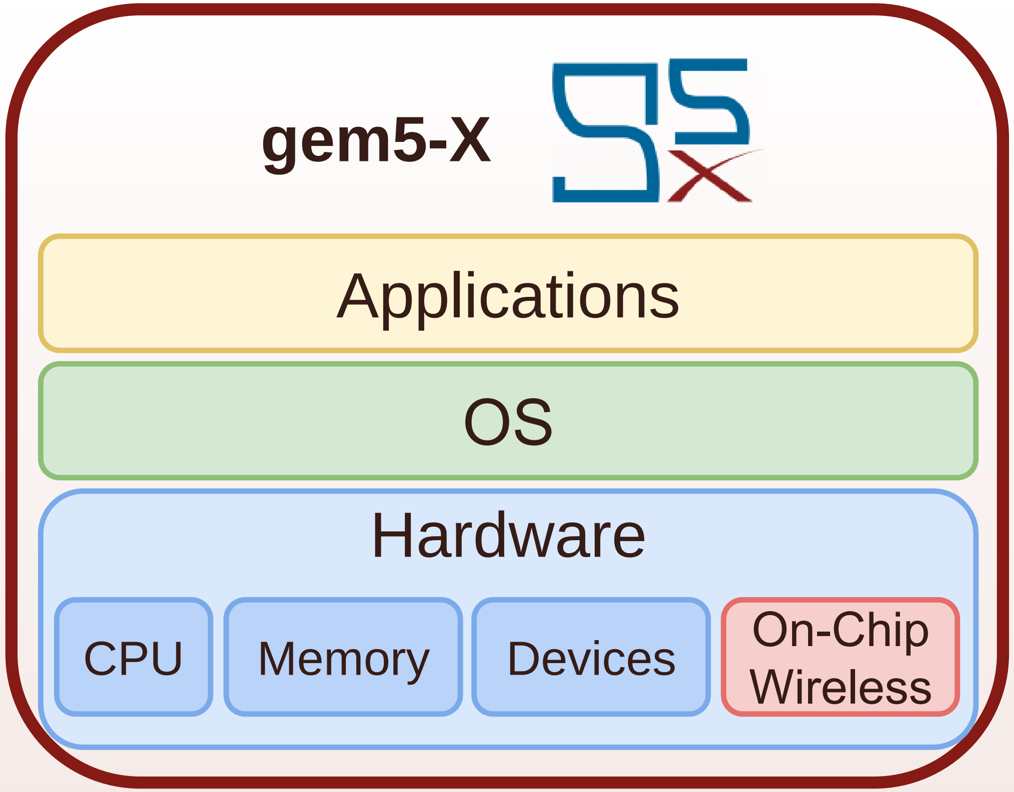
- Assess impact of wireless interconnects on the system

↓

Full System Simulation

  - Hardware and software

- **gem5-X [3]**
  - Extensible with new architectures
  - Model of on-chip wireless links



gem5-X

Applications

OS

Hardware

CPU

Memory

Devices

On-Chip Wireless

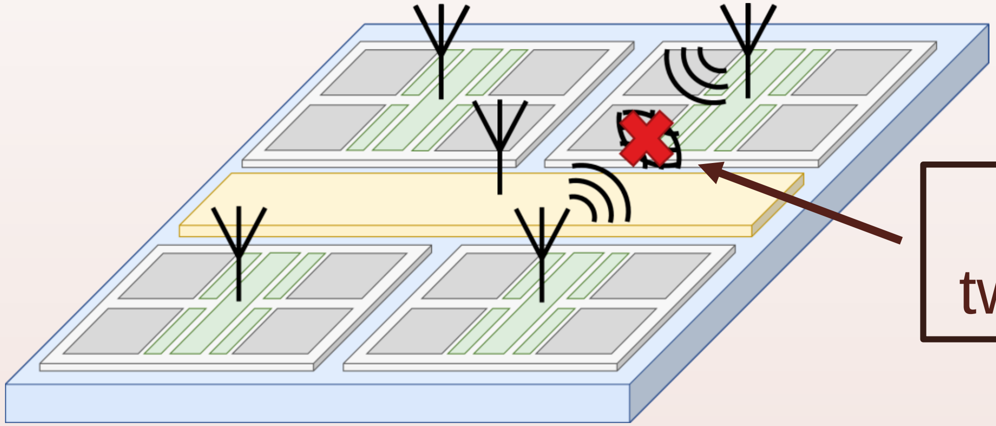
[3] Y. Qureshi et al., TACO, 2021

Exploration of Wireless Configurations

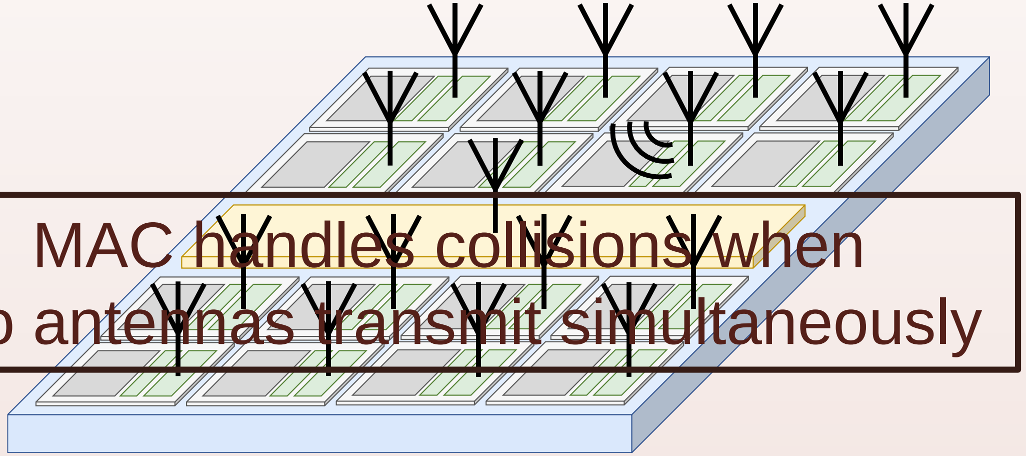
- System integration
  - Which components are connected by wireless links

- Configuration of the transceivers
  - Channel bandwidth
  - Communication latency
  - Medium Access Control (MAC)

One transceiver per cluster



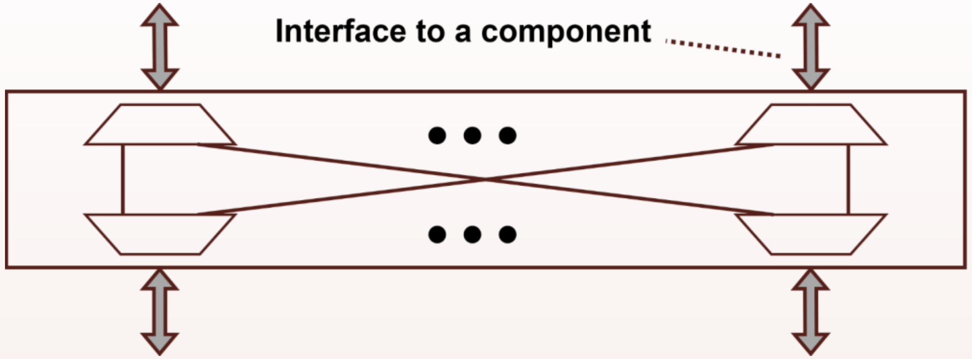
One transceiver per CPU



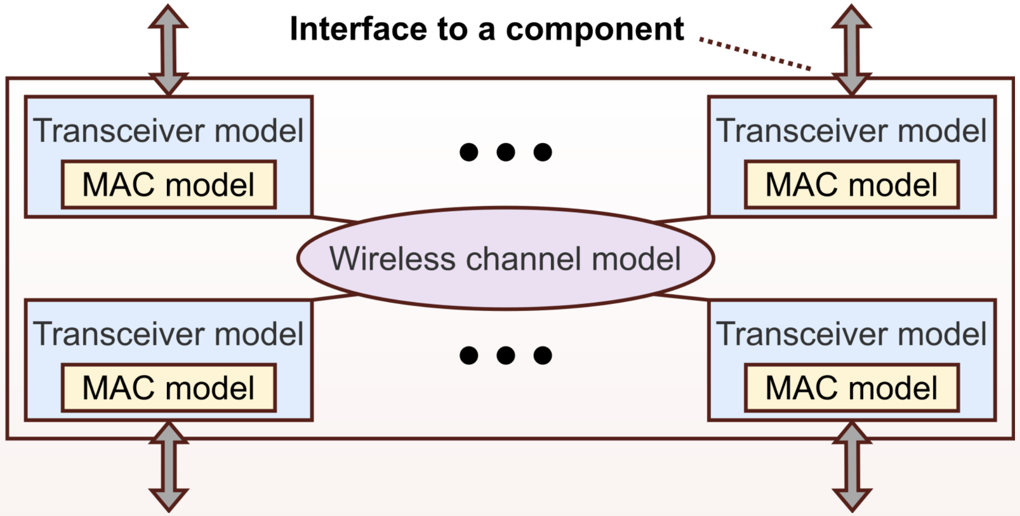
MAC handles collisions when two antennas transmit simultaneously

Modelling Methodology

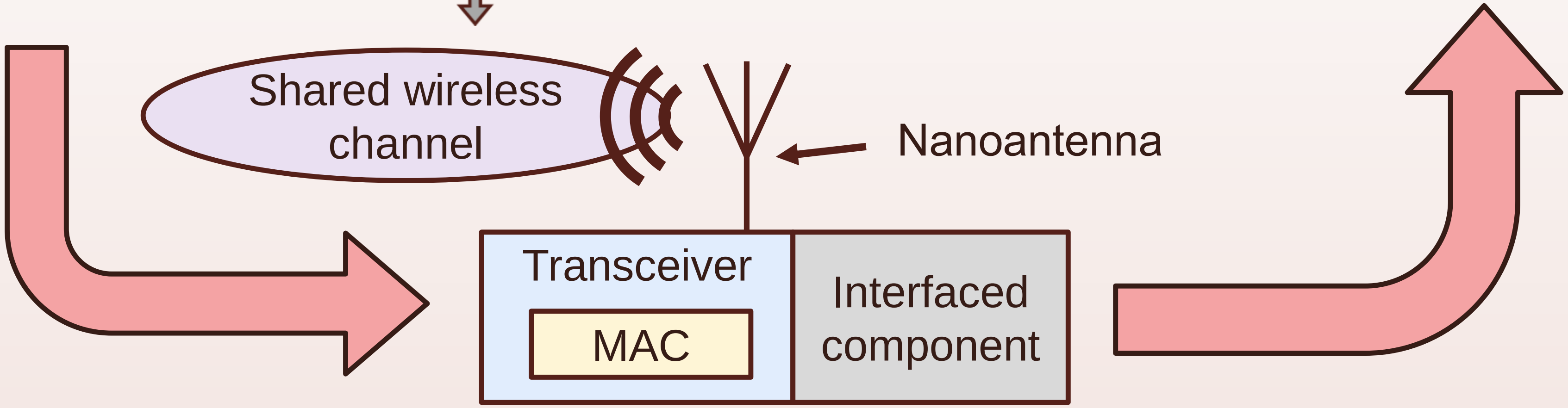
- Model wireless chiplet interconnect
- Start from gem5 interconnect model
  - **Gem5 crossbar**



gem5-X wireless crossbar

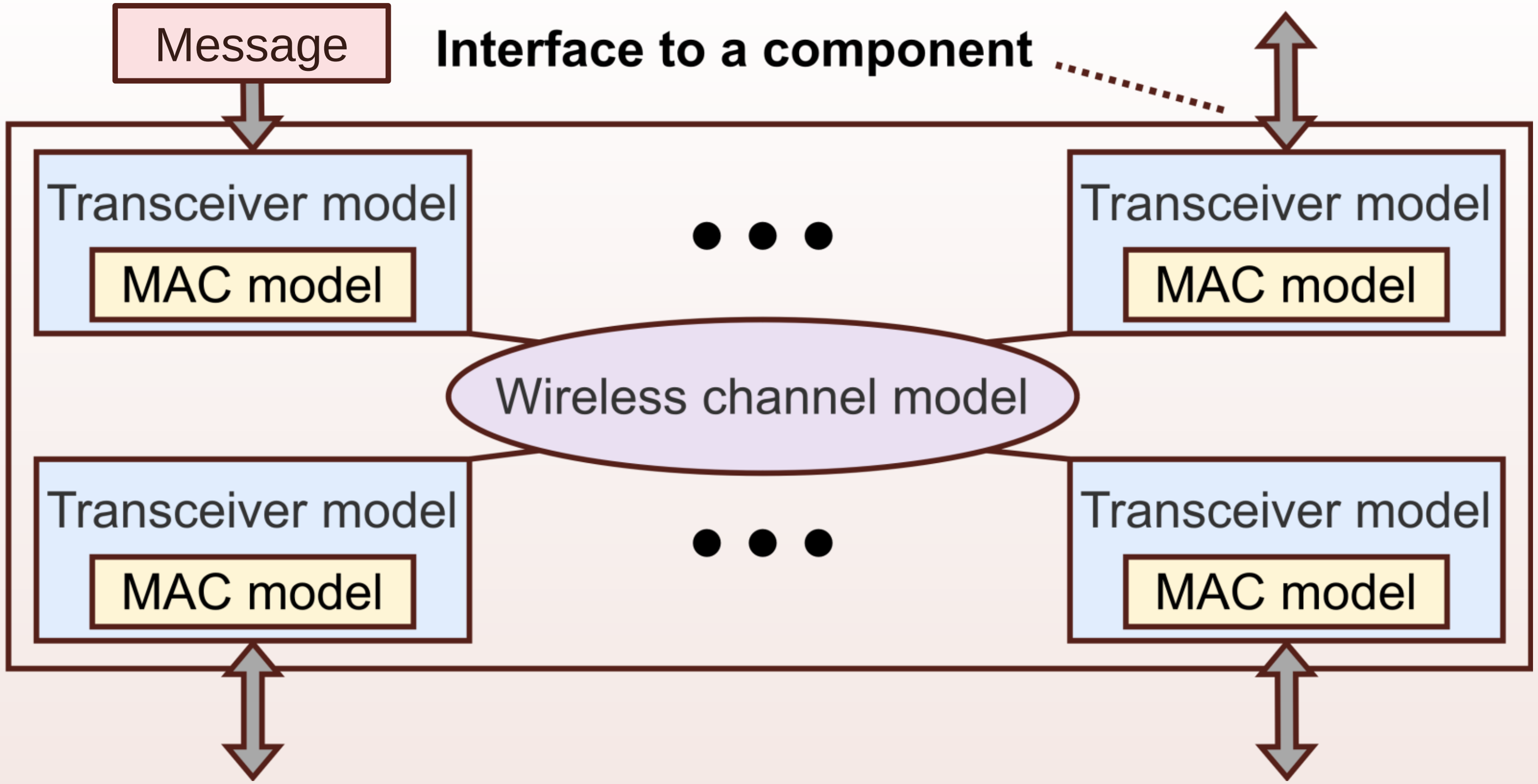


Add wireless elements

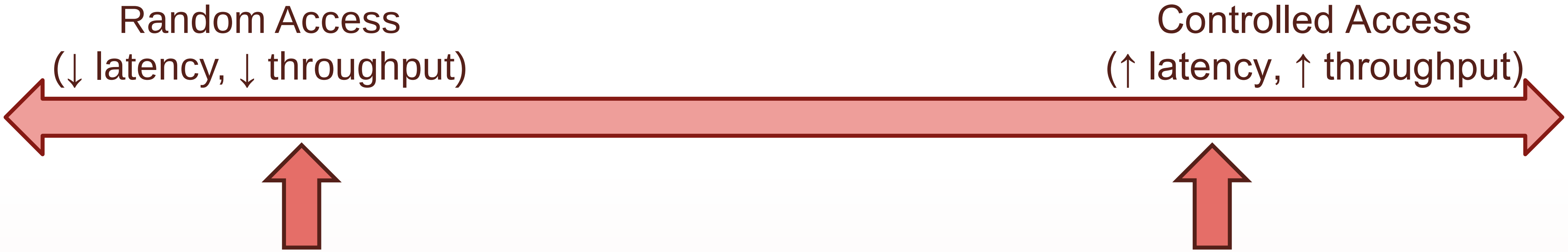


Wireless Communication Process

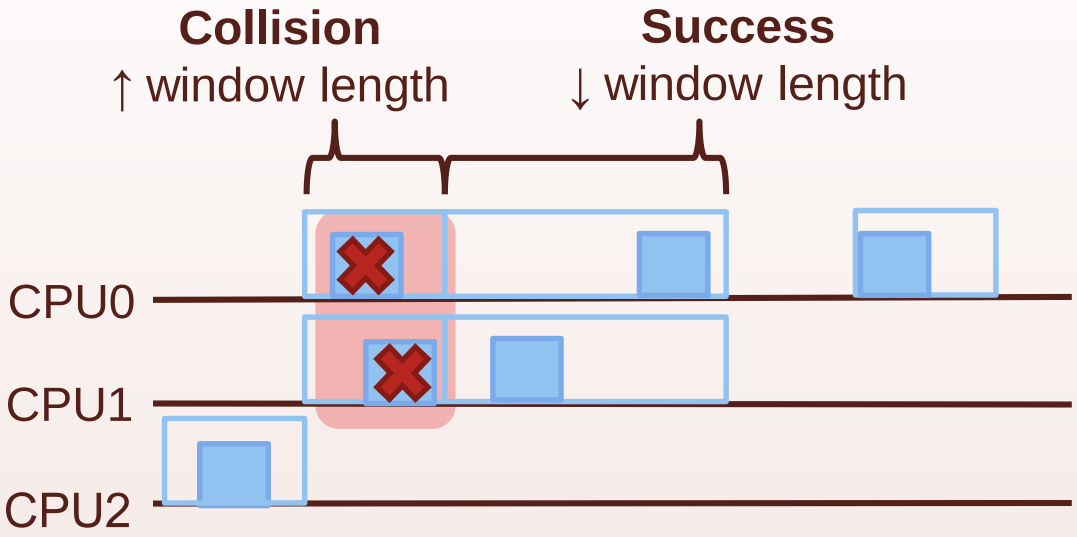
1. Transmission request
2. MAC protocol
3. Apply delays
4. Transmission successful



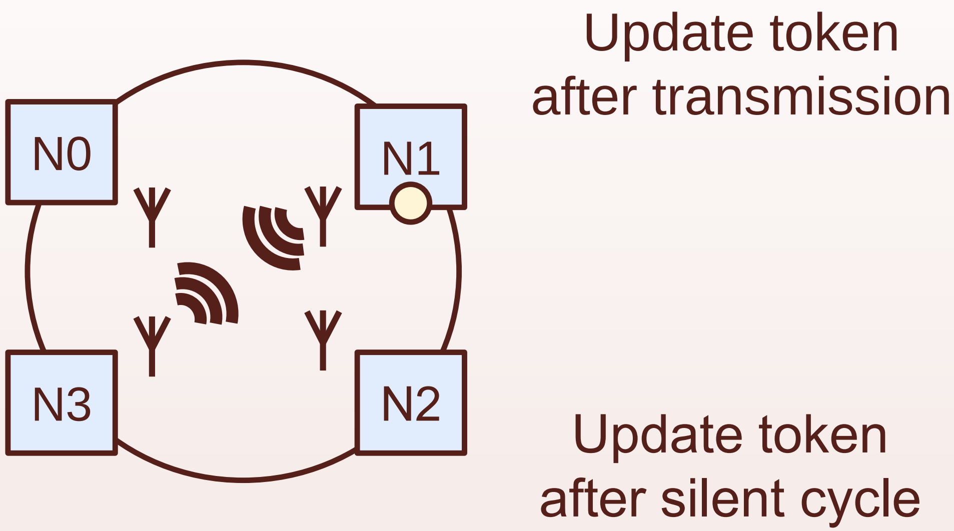
Modelling MAC Protocols



Exponential backoff



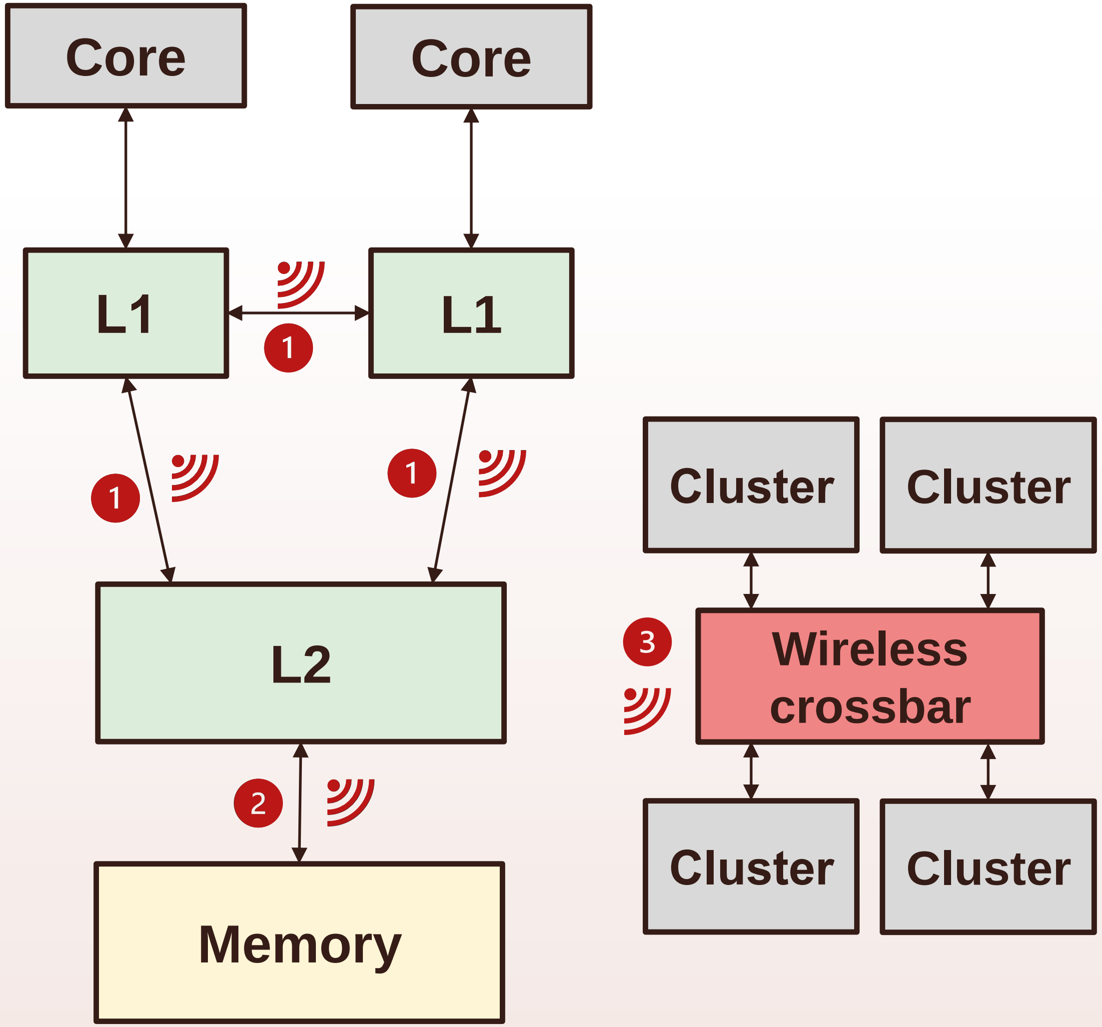
Token passing



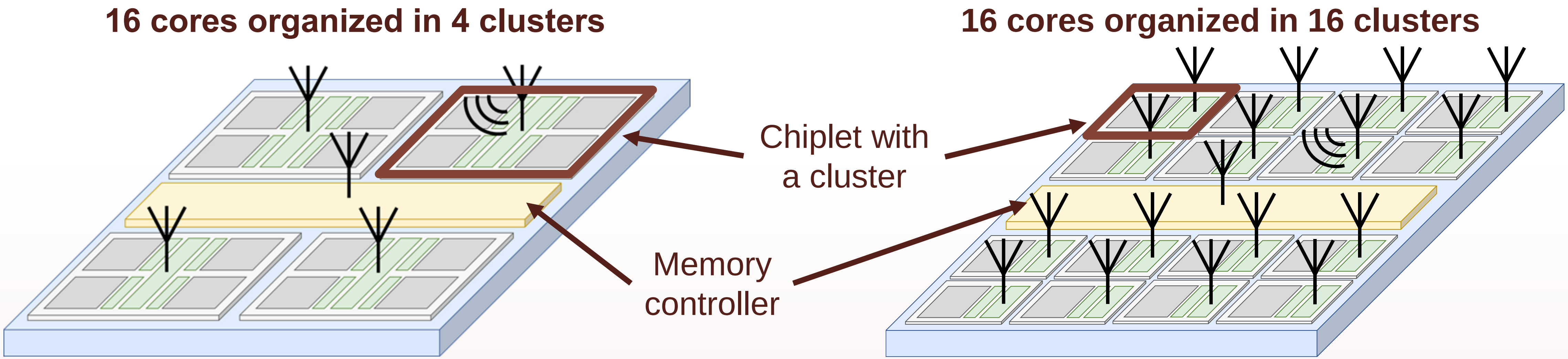
Code available at *https://github.com/gem5-X/On-Chip-Wireless*

Integrating Wireless Interconnects in gem5-X

- Simple integration
- Unmodified OS and applications
- Emulation of interconnect options
  1. Connection between L1s and L2
  2. Memory bus
  3. Connection between clusters



Explored Architectures


16 cores organized in 4 clusters

16 cores organized in 16 clusters

Chiplet with a cluster

Memory controller

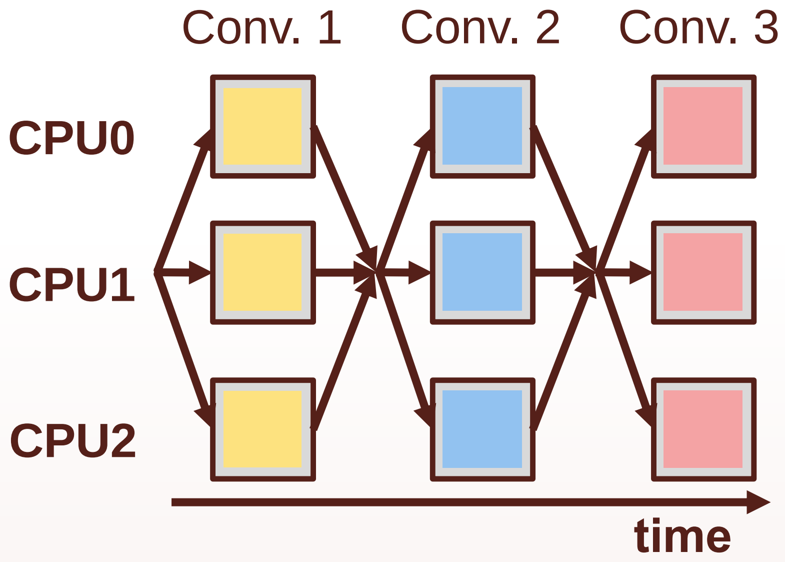
Processors		16x Out-of-Order ARMv8 cores @ 2 GHz
L1-I Caches		16x private, 32 kB, 2-way, 2 cycle access
L1-D Caches		16x private, 32 kB, 2-way, 2 cycle access
L2 Caches	**4 clusters**	4x shared, 1 MB, 2-way, 20 cycle access
	16 clusters	16x private, 256 kB, 2-way, 20 cycle access
Memory		DDR4 2400 MHz, 4 GB
OS		Ubuntu LTS 16.04

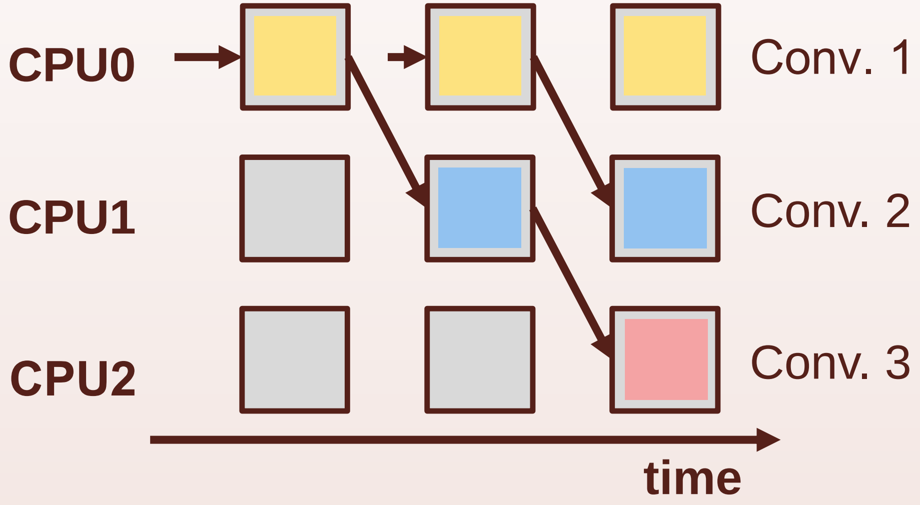
Explored Workloads

- Communication-intensive benchmarks
  - Assess interconnect under heavy traffic
  - STREAM [4], SPLASH-2 [5]
- Neural Networks [6, 7, 8]
  - Bursty traffic
  - Mapping

Intra-layer

Inter-layer

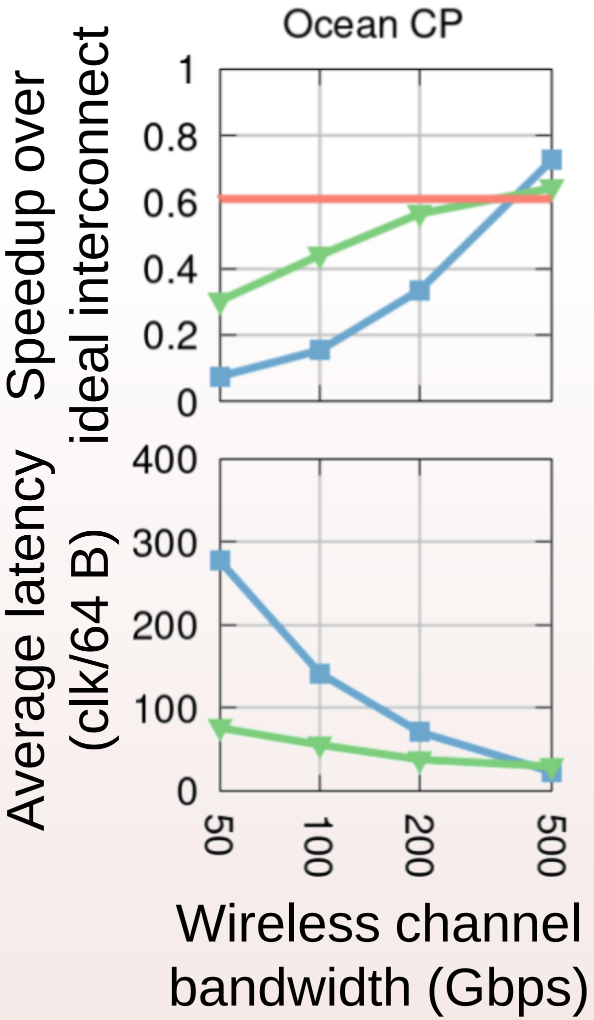




[4] J. McCalpin, TCCA, 1995
[5] S. Woo et al., ISCA, 1995

[6] K. Chatfield et al., ArXiv, 2014
[7] A. Howard et al., ArXiv, 2017
[8] M. Sandler et al., ArXiv, 2018

4 Clusters Results



Baseline: ideal interconnect

- Transfers take one cycle

Wired chiplet interconnect

- State-of-the-art, UCIe compliant
- 112 Gbps, 100 clk latency [9, 10]

Wireless with exponential backoff

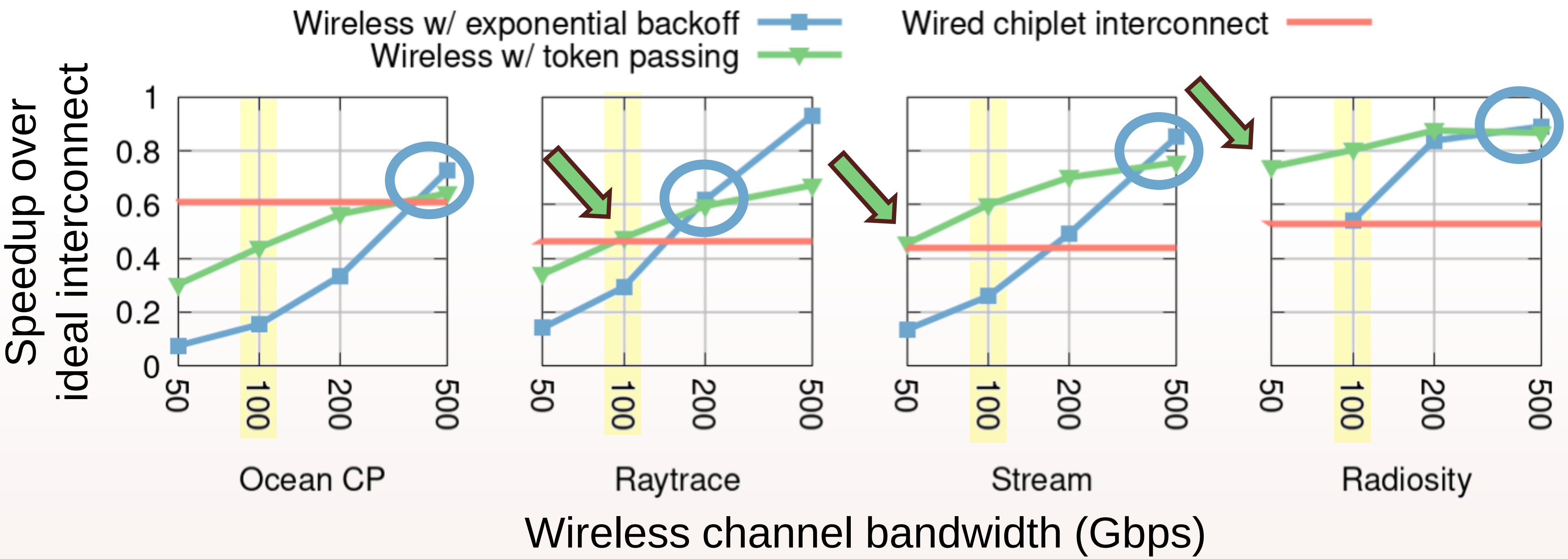
Wireless with token passing

Exponential backoff latency depends on bandwidth → collisions

Token passing keeps latency low across bandwidths

[9] S. Naffziger et al., ISCA, 2021 [10] Synopsis, “Die-to-Die IP Solutions”, 2022

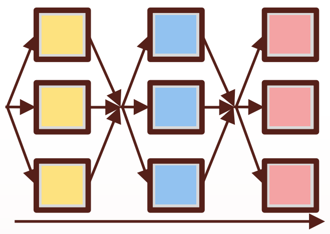
4 Clusters: Intensive Benchmarks



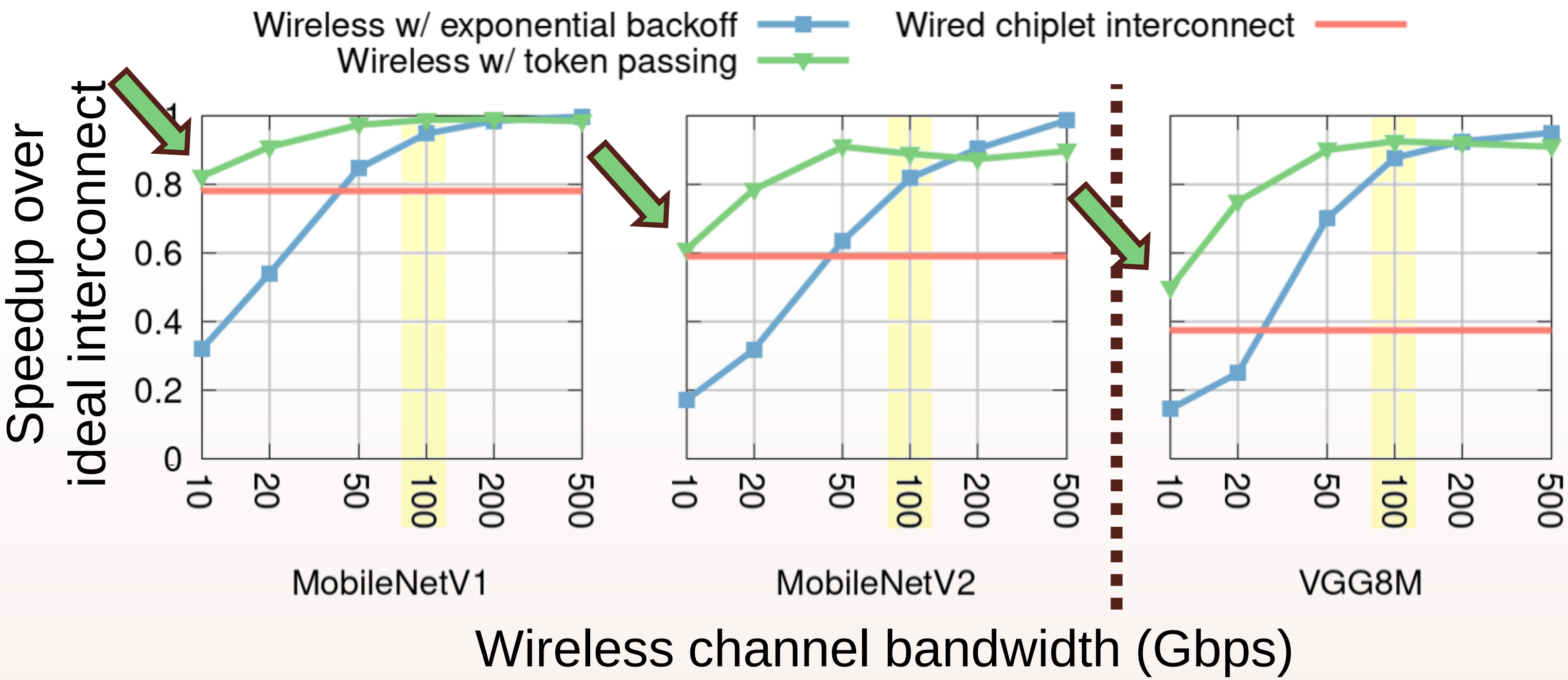
Token passing outperforms wired interconnects for most workloads at 100 Gbps

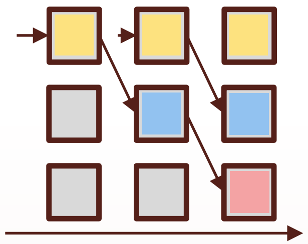
Exponential backoff needs high bandwidths to surpass token passing

4 Clusters: Neural Networks



Intra-layer mapping





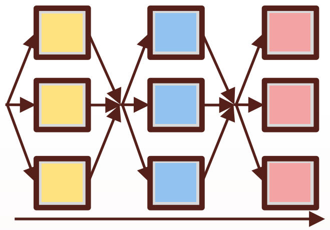
Inter-layer mapping

Token passing is better than wired from 10 Gbps

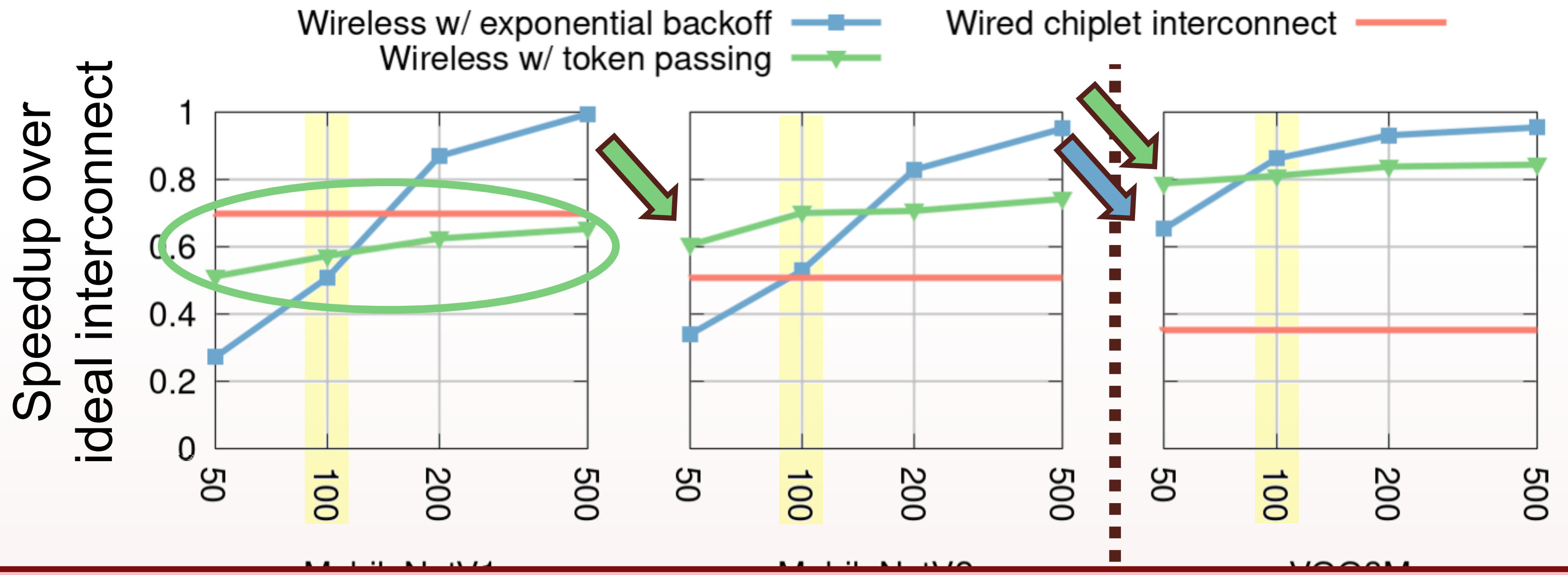
Token passing is better
Many simultaneous transmissions

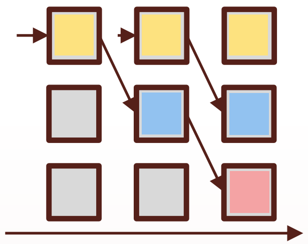
Exponential backoff is better
Transmission spread in time

16 Clusters: Neural Networks



Intra-layer mapping





Inter-layer mapping

Wireless in many-node networks can support bursty workloads

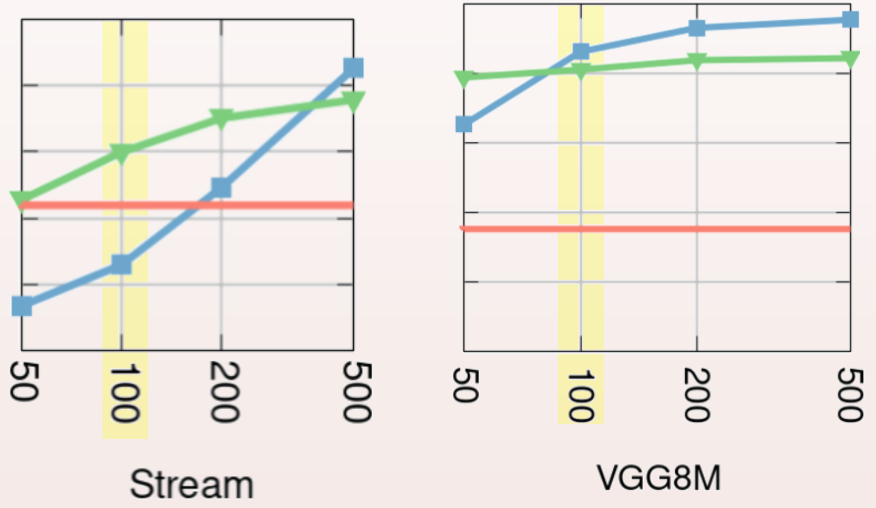
- More nodes → difficulties with high traffic loads
- Still enough for lower bandwidth requirements

Exponential backoff is good for many-node inter-layer

Conclusions

- System-level model of in-package wireless communication
  - Code available now!
- Wireless competitive with wired chiplet interconnects
  - Token passing works well at demonstrated bandwidths

https://github.com/gem5-X/On-Chip-Wireless



System-Level Exploration of In-Package Wireless Communication for Multi-Chiplet Platforms

Thank you!

Rafael Medina, Joshua Klein, Giovanni Ansaloni, Marina Zapater,
Sergi Abadal, Eduard Alarcón, David Atienza

Correspondence: rafael.medinamorillas@epfl.ch

EMBEDDED SYSTEMS LABORATORY

gem5-X

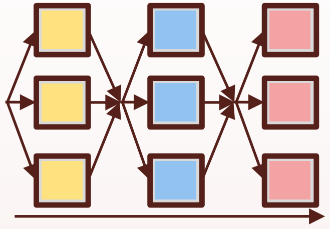
- Available for download at ESL repository
  - https://github.com/gem5-X
- Documentation and manual
  - https://www.epfl.ch/labs/esl/research/2d-3d-system-on-chip/gem5-x/

On-Chip Wireless in gem5-X

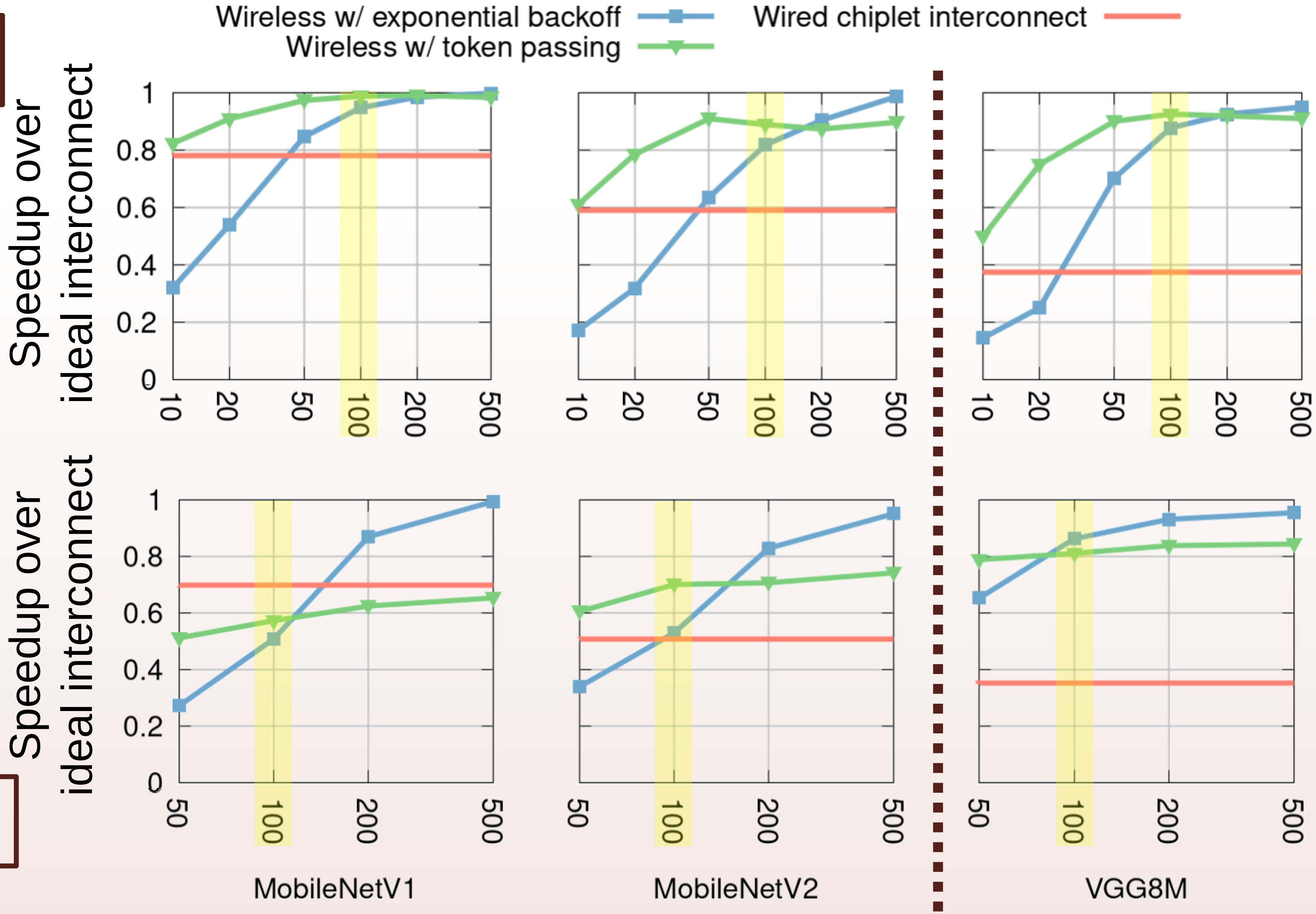
- Available for download at ESL repository
  - https://github.com/gem5-X/On-Chip-Wireless
- Documentation and manual
  - https://www.epfl.ch/labs/esl/research/2d-3d-system-c

4 Clusters vs. 16 Clusters

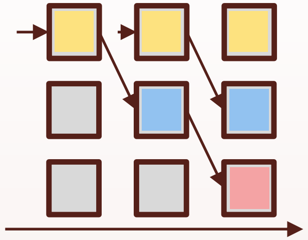
4 Clusters



Intra-layer parallelism

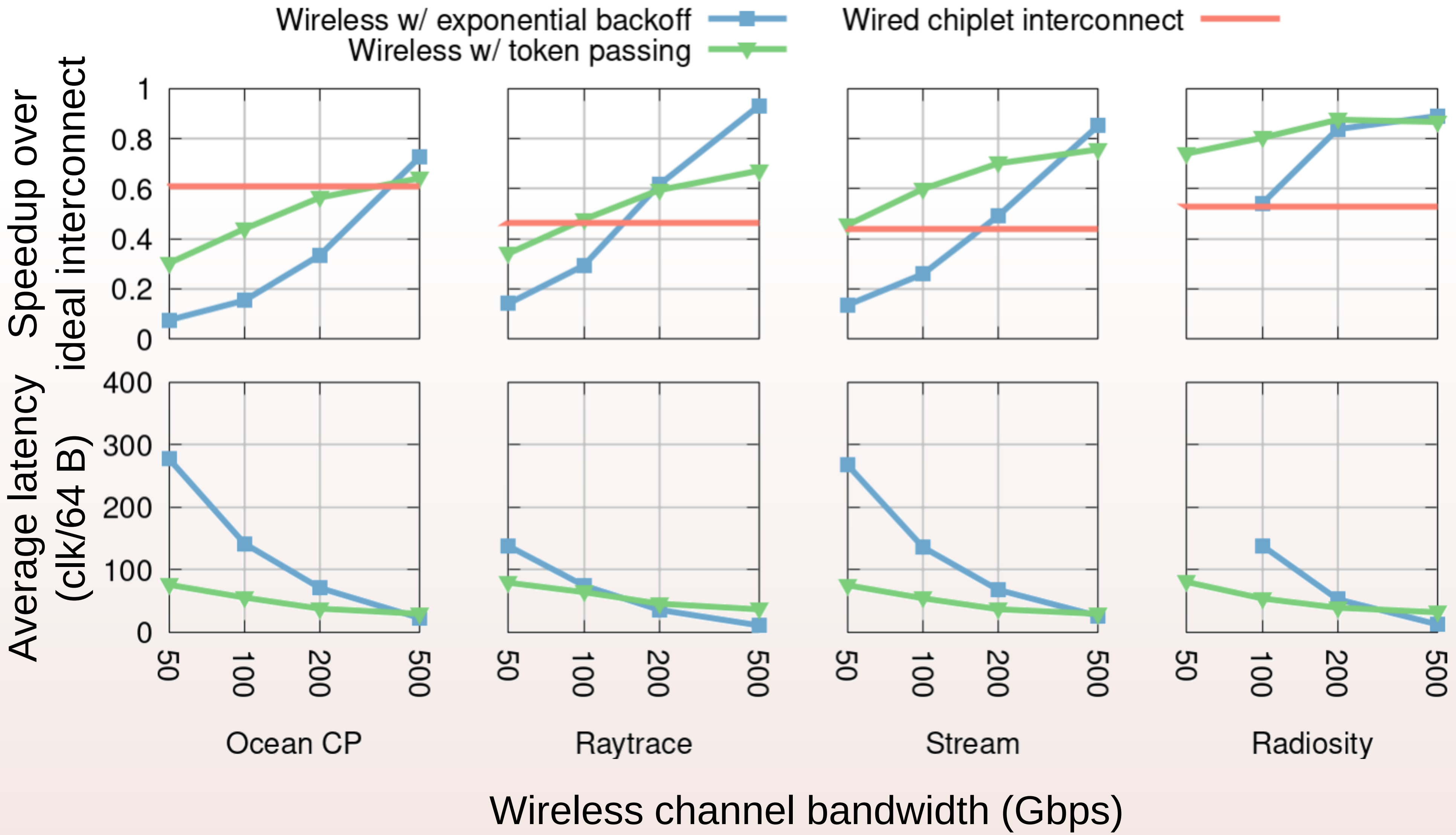


16 Clusters

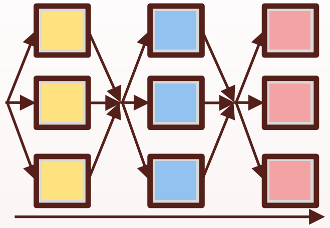


Inter-layer parallelism

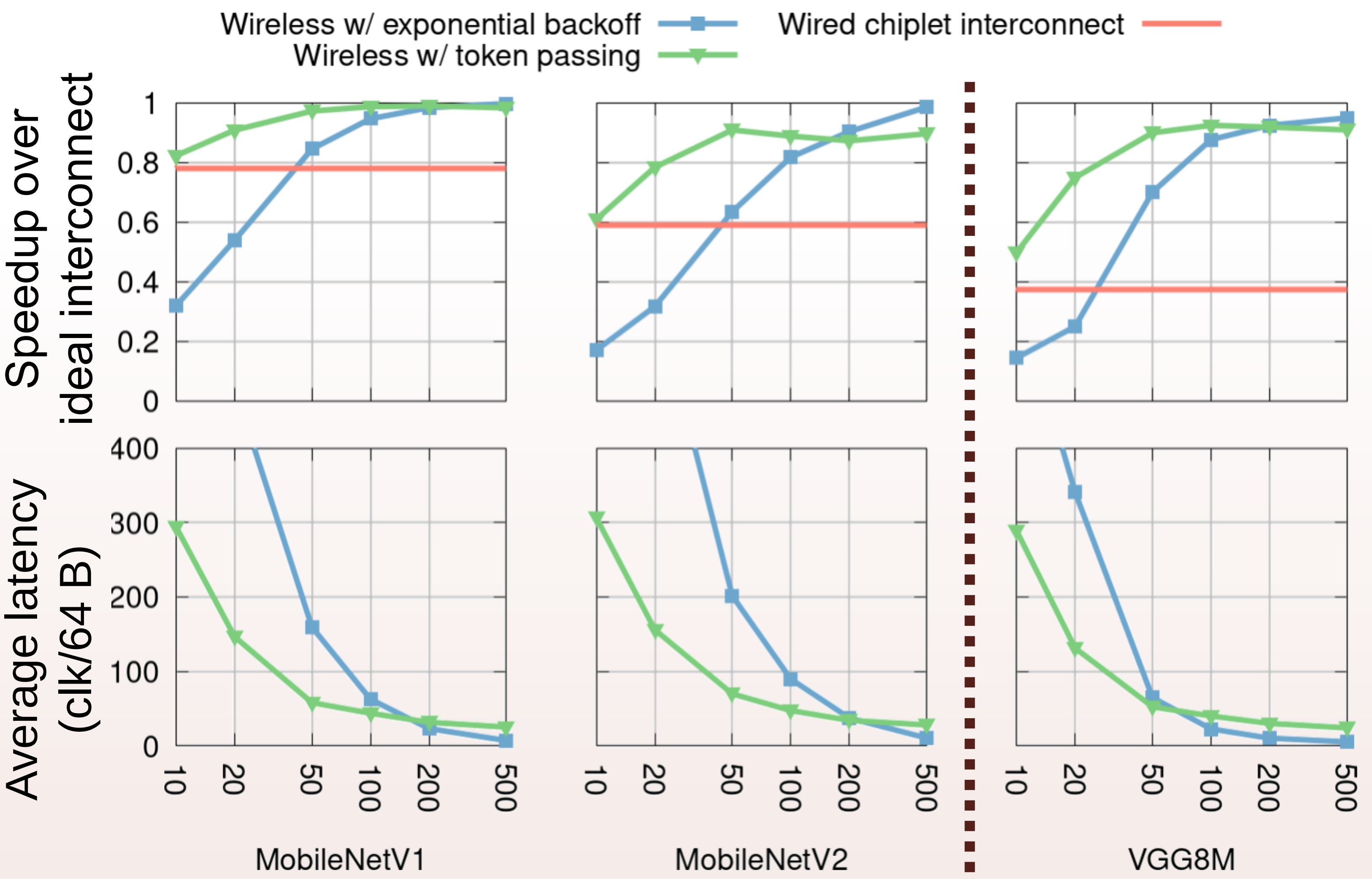
4 Clusters: Intensive Benchmarks



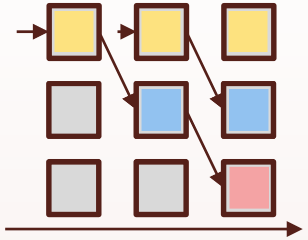
4 Clusters: CNNs



Intra-layer parallelism

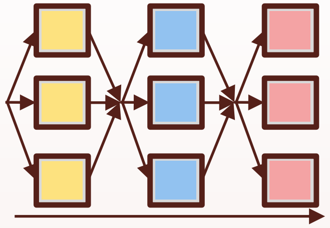


Wireless channel bandwidth (Gbps)

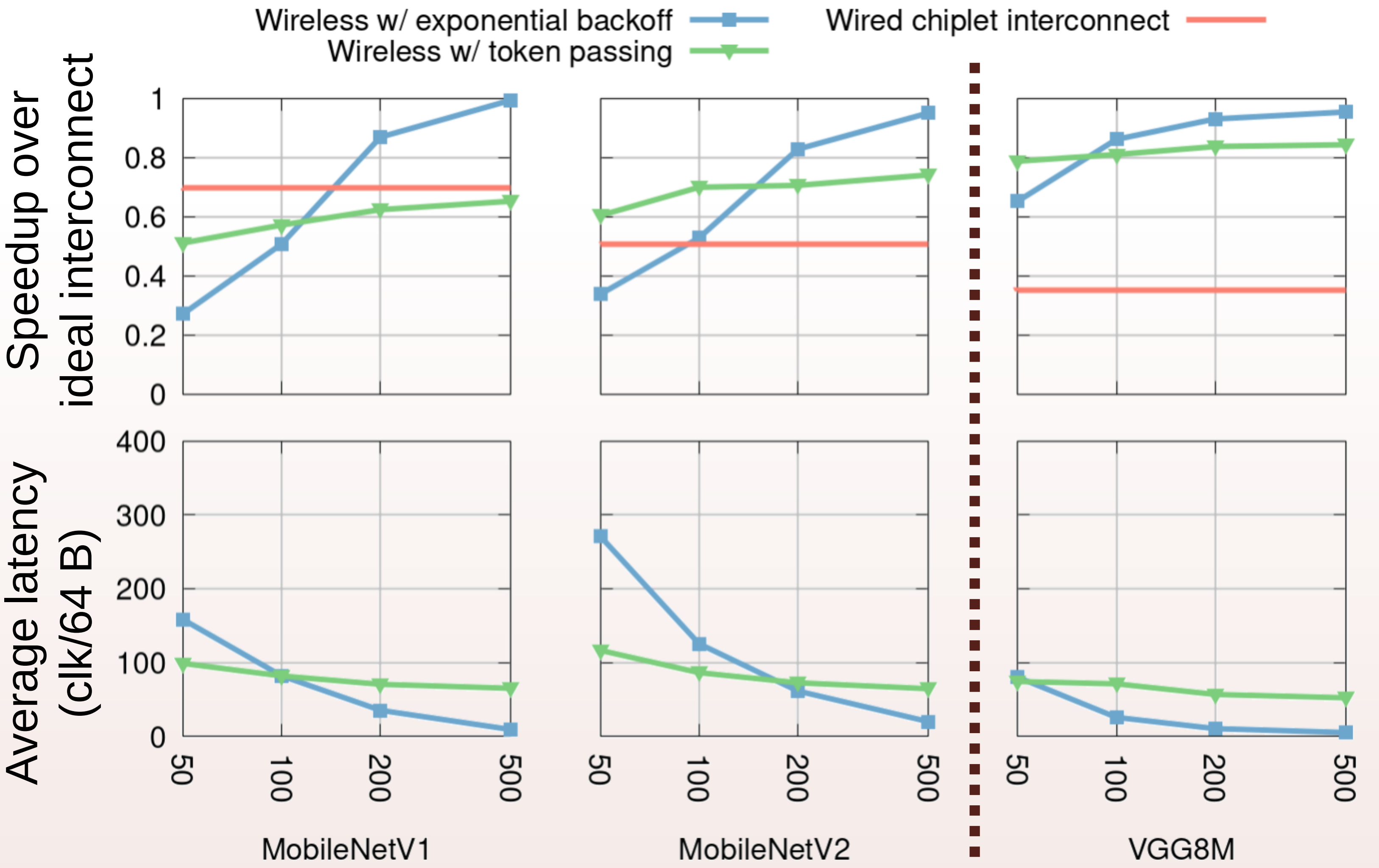


Inter-layer parallelism

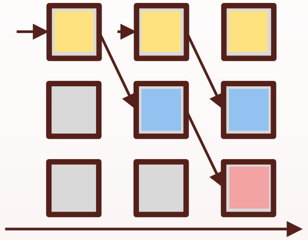
16 Clusters: CNNs



Intra-layer parallelism



Wireless channel bandwidth (Gbps)



Inter-layer parallelism