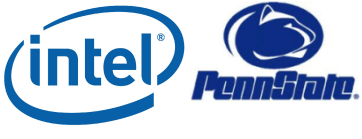
A Heterogeneous Multiple Network-On-Chip Design: An Application-Aware Approach

Asit K. Mishra

Onur Mutlu

Chita R. Das



Executive summary

• **Problem**: Current day NoC designs are **agnostic** to application requirements and are provisioned for the general case or worst case. Applications have widely differing demands from the network

• **Our goal**: To design a NoC that can satisfy the **diverse** dynamic performance requirements of applications

• **Observation**: Applications can be divided into two general classes in terms of their requirements from the network: **bandwidth-sensitive** and **latency-sensitive**
  - Not all applications are equally sensitive to bandwidth and latency

• **Key idea**: Design two NoC
  - Each sub-network customized for either BW or LAT sensitive applications
  - Propose **metrics** to classify applications as BW or LAT sensitive
  - Prioritize applications' packets within the sub-networks based on their sensitivity

• **Network design**: BW optimized network has wider link width but operates at a lower frequency and LAT optimized network has narrow link width but operates at a higher frequency

• **Results**: Our proposal is significantly better when compared to an iso-resource monolithic network (5%/3% weighted/instruction throughput improvement and 31% energy reduction)

Resource requirements of various applications - I

Impact of channel bandwidth on application performance

- Channel bandwidth affects network latency, throughput and energy/power

- Increase in channel BW leads to
  - Reduction in packet serialization
  - Increase in router power

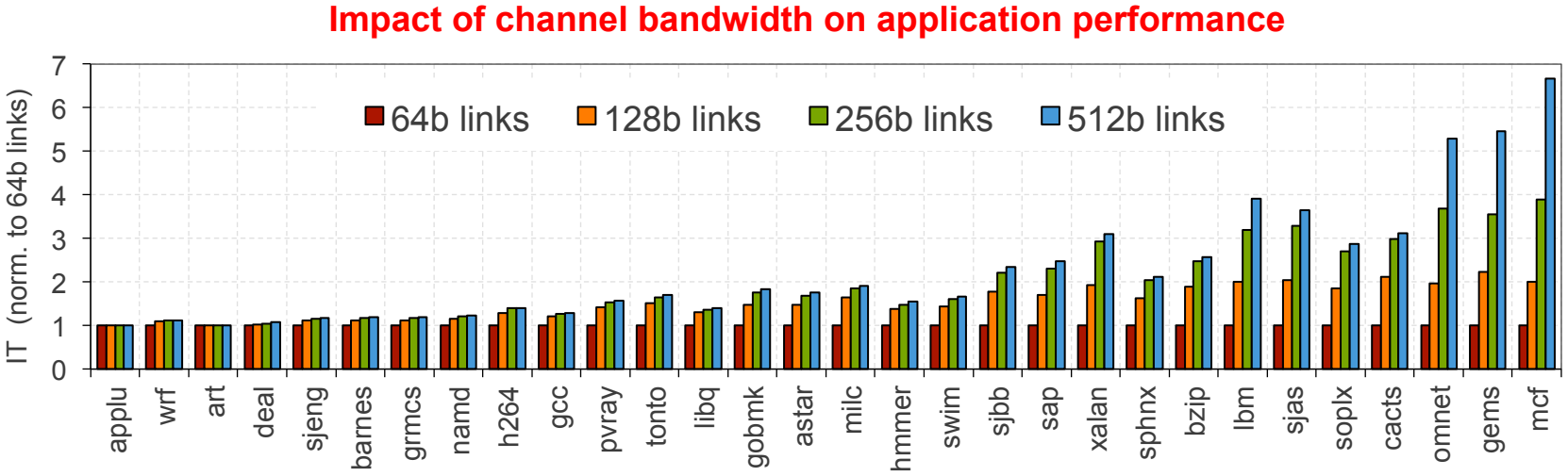
Resource requirements of various applications - I

Impact of channel bandwidth on application performance

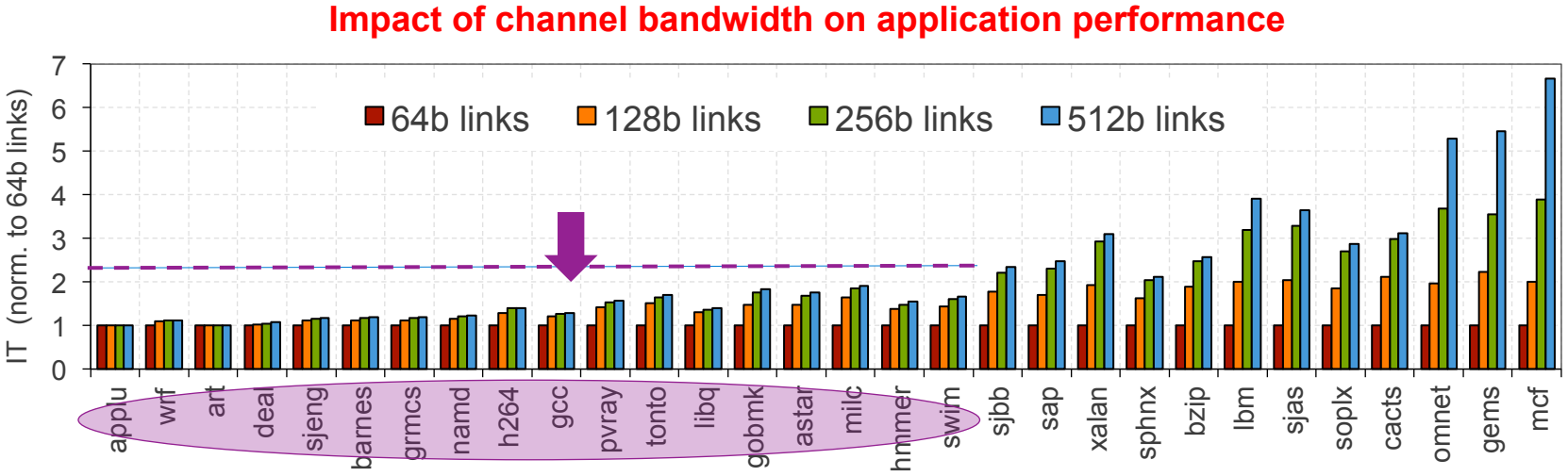
Simulation settings:

- 8x8 multi-hop packet based mesh network
- Each node in the network has an OoO processor (2GHz), private L1 cache and a router (2GHz)
- Shared 1MB per core shared L2
- 6VC/PC, 2 stage router

Resource requirements of various applications - I

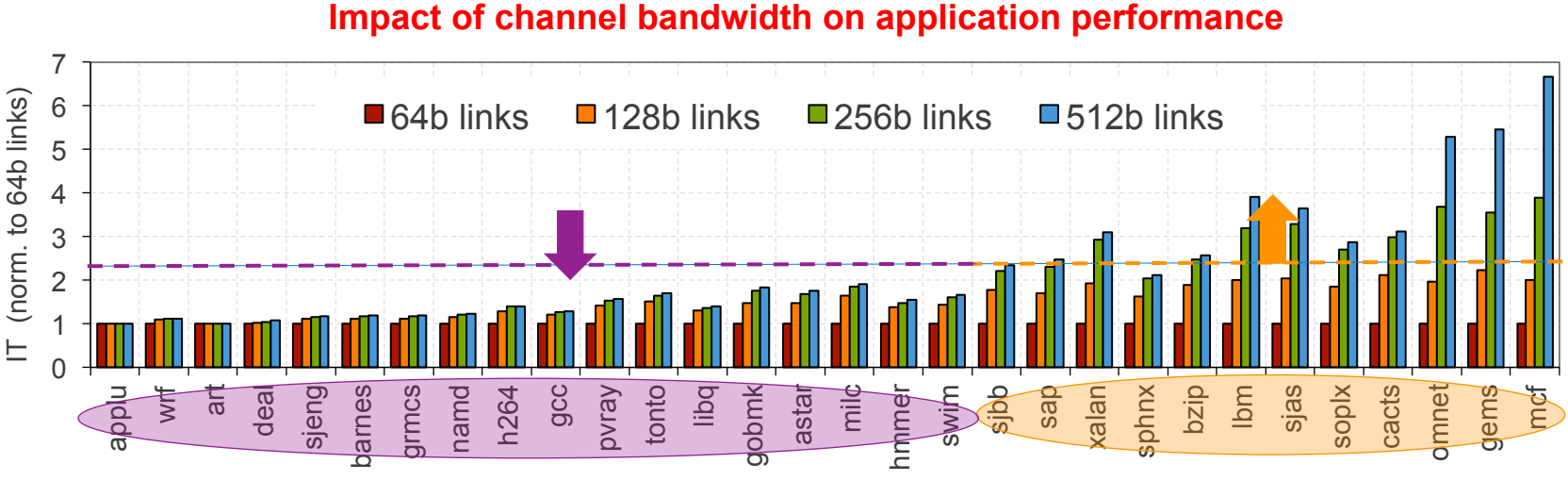


Resource requirements of various applications - I



1. 18/30 (21/36 total) applications' performance is agnostic to channel BW (8x BW inc. → less than 2x performance inc.)

Resource requirements of various applications - I



1. 18/30 (21/36 total) applications' performance is agnostic to channel BW (8x BW inc. → less than 2x performance inc.)

2. 12/30 (15/36 total) applications' performance scale with increase in channel BW (8x BW inc. → at least 2x performance inc.)

Resource requirements of various applications - II

Impact of network latency on application performance

- Reduction in router latency (by increasing frequency)
  - Reduction in packet latency
  - Increase in router power consumption

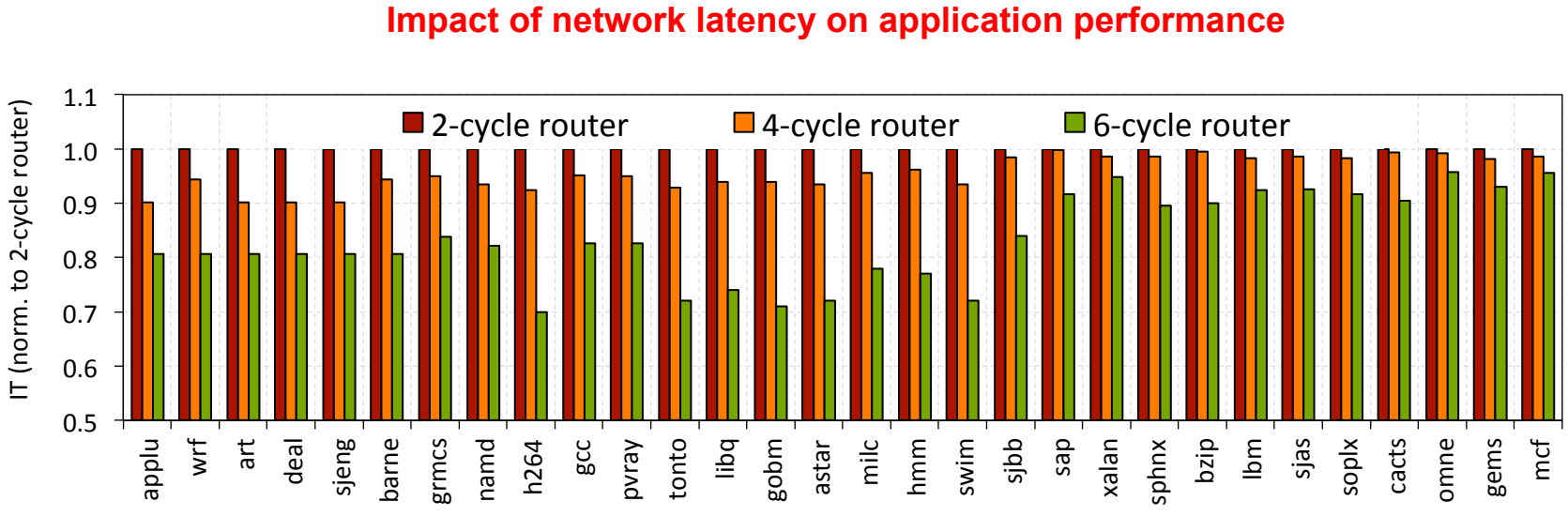
Resource requirements of various applications - II

Impact of network latency on application performance

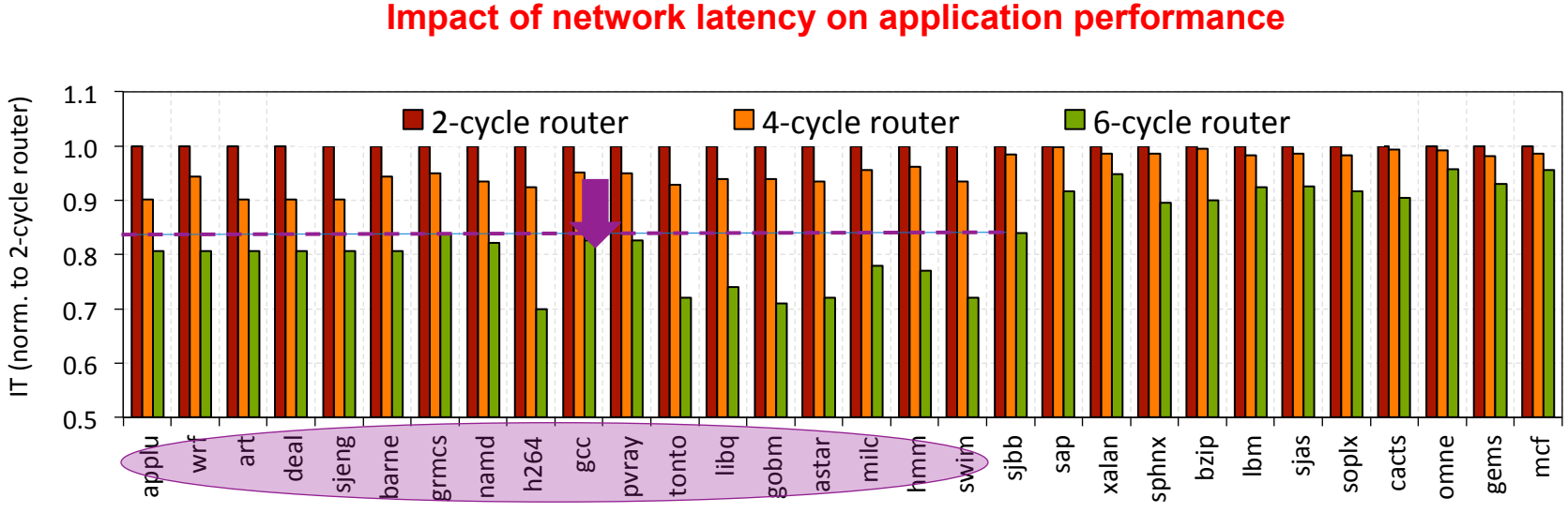
Simulation settings:

- … same as last experiment
- 128b links
- Added dummy stages (2-cycle and 4-cycle) to each router

Resource requirements of various applications - II

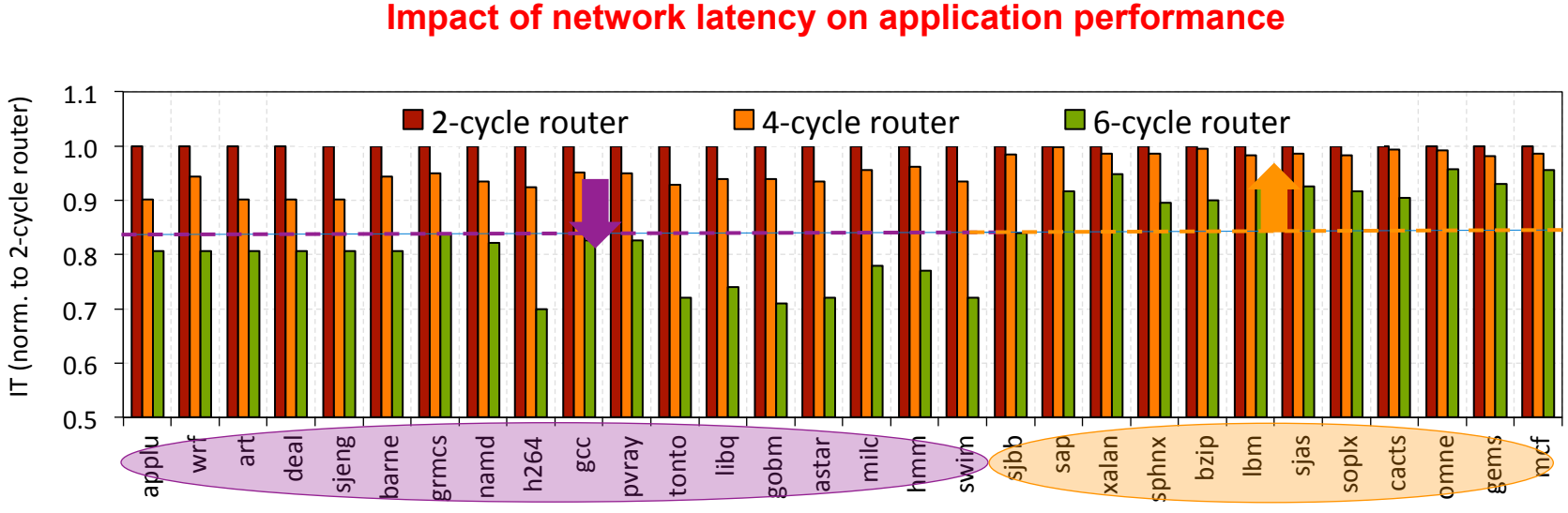


Resource requirements of various applications - II



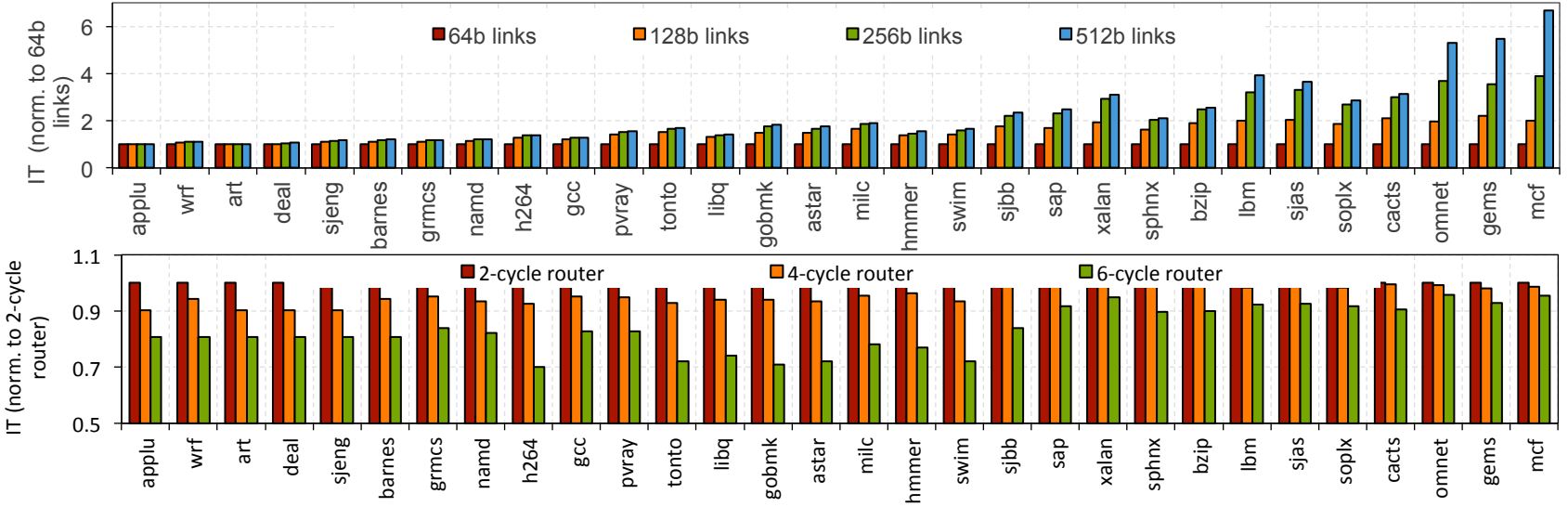
1. 18/30 (21/36 total) applications' performance is sensitive to network latency (3x latency reduction → at least 25% performance improvement)

Resource requirements of various applications - II

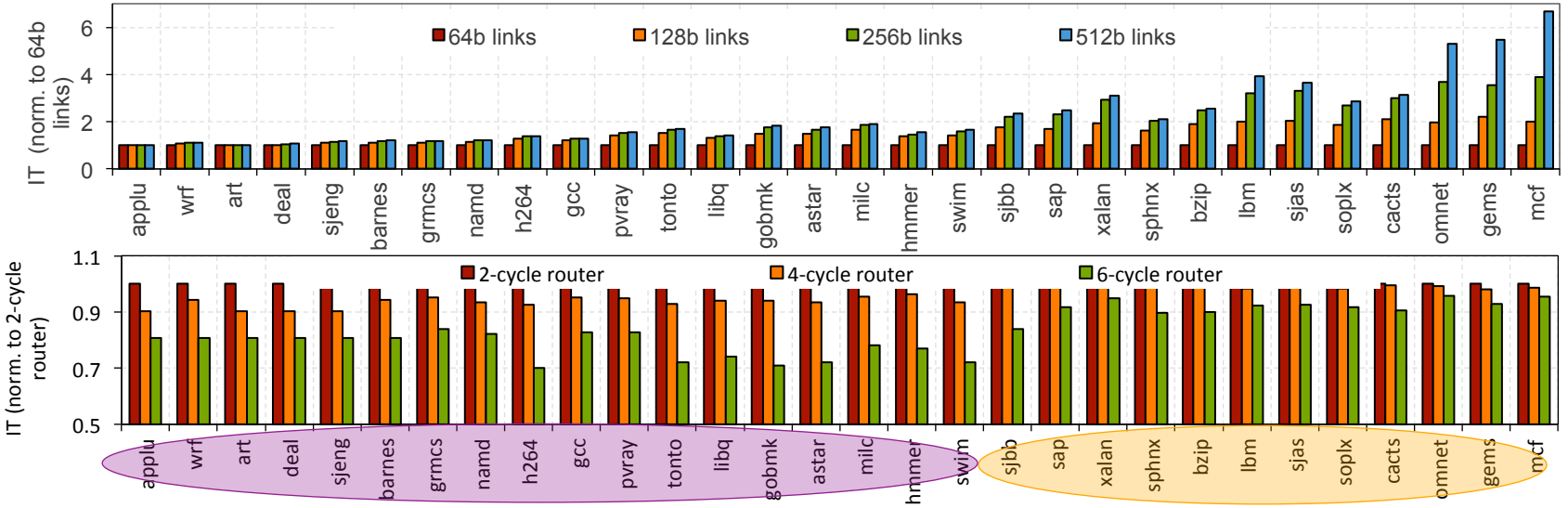


1. 18/30 (21/36 total) applications' performance is sensitive to network latency (3x latency reduction → at least 25% performance improvement)
2. 12/30 (15/36 total) applications' performance is marginally sensitive to network latency (3x latency increase → less than 15% performance improvement)

Application-aware approach to designing multiple NoCs



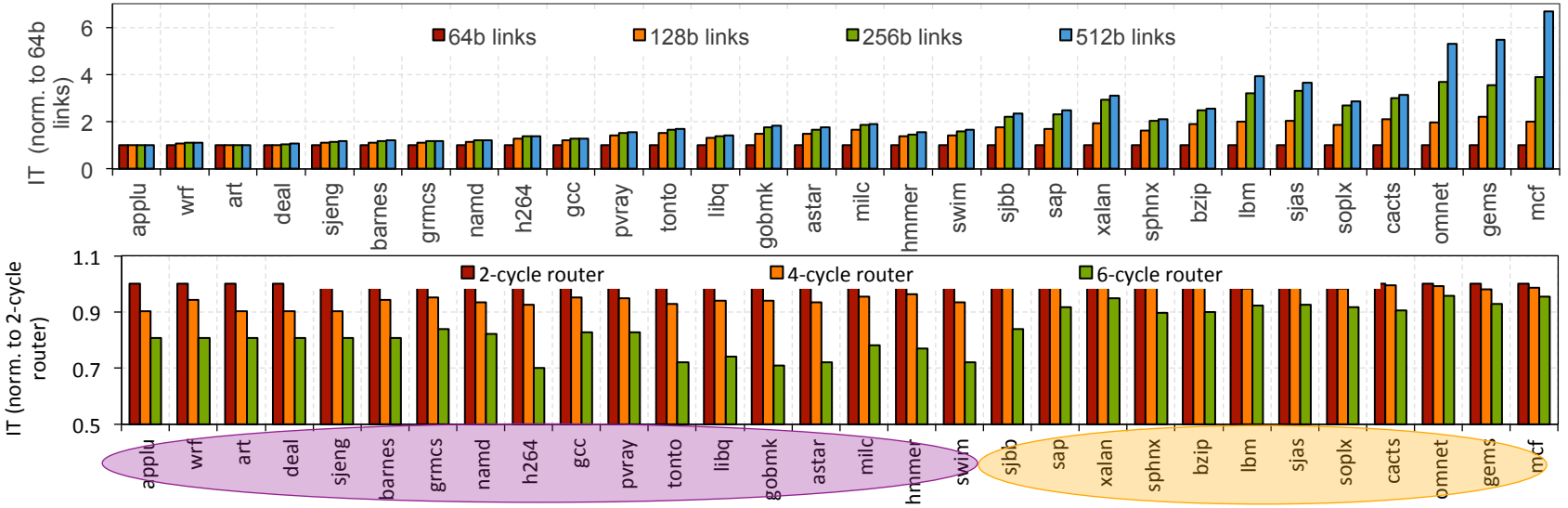
Application-aware approach to designing multiple NoCs



Based on the observations:

1. Applications can be classified into distinct classes: typically LAT/BW sensitive
2. LAT sensitive applications can benefit from low network latency
3. BW sensitive applications can benefit from high network bandwidth
4. Not all applications are equally sensitive to either LAT or BW
5. Monolithic network cannot optimize both classes simultaneously

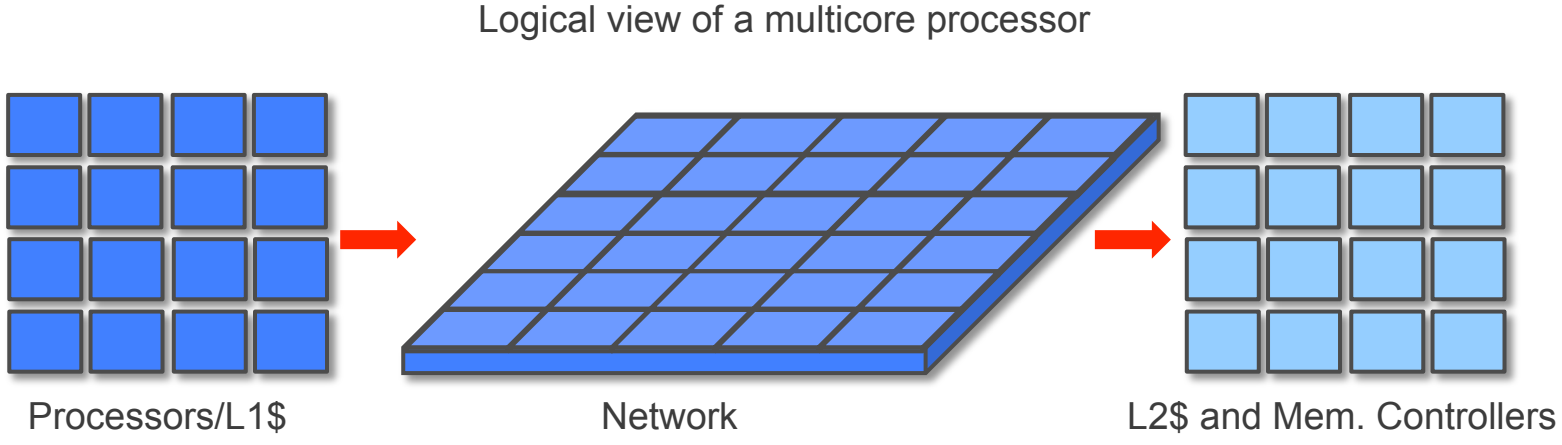
Application-aware approach to designing multiple NoCs



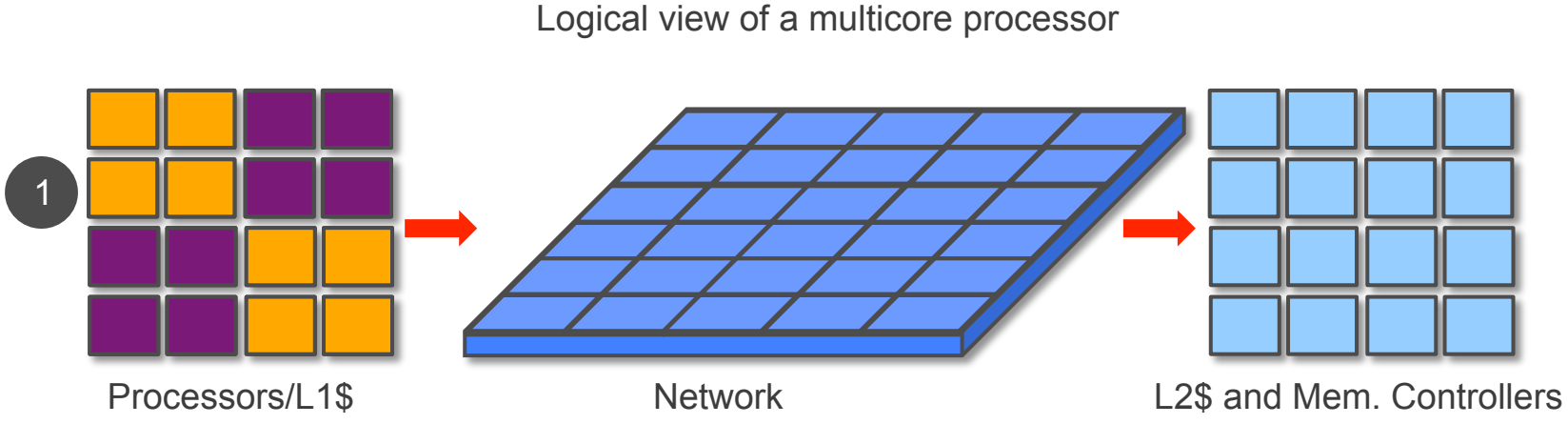
Solution

Two NoCs where each (sub)network is optimized for either LAT or BW sensitive applications

Design methodology

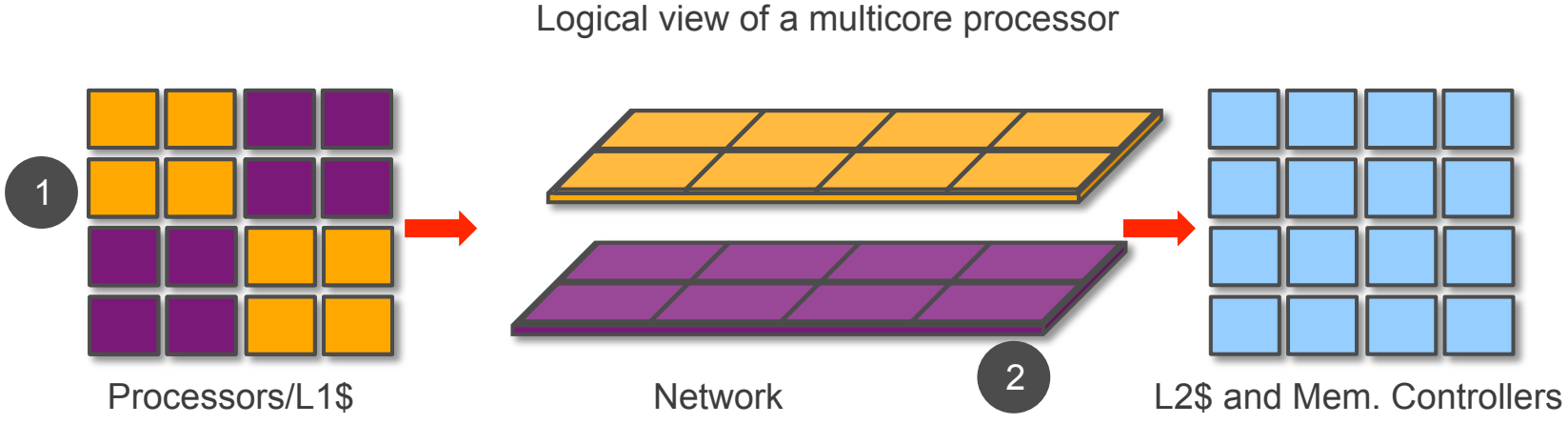


Design methodology



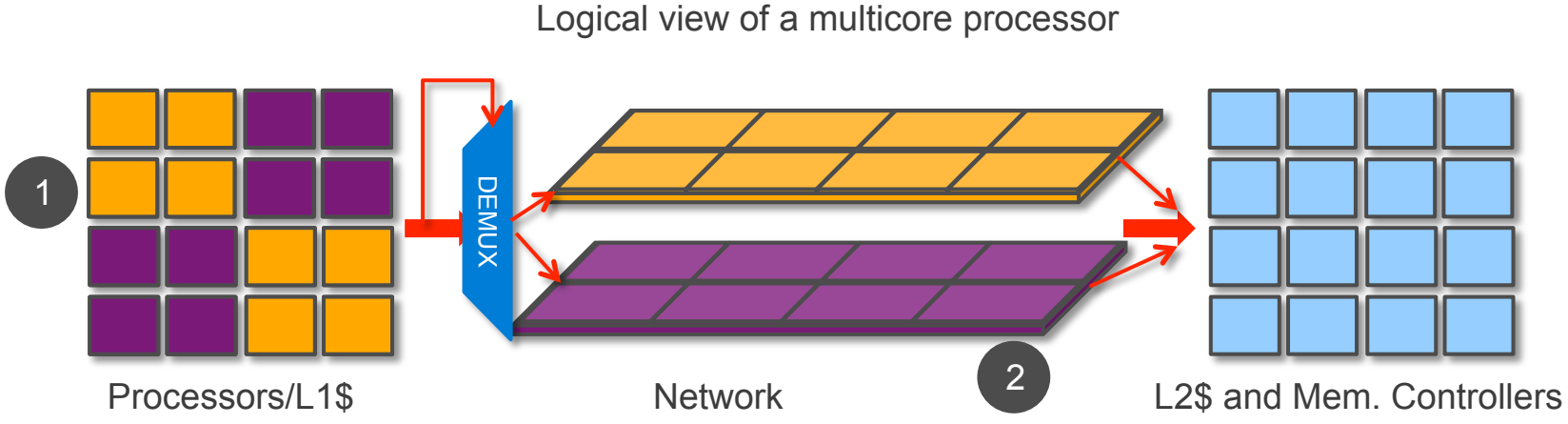
1 Identify LAT/BW sensitive applications
- Proposes a novel dynamic application classification scheme

Design methodology



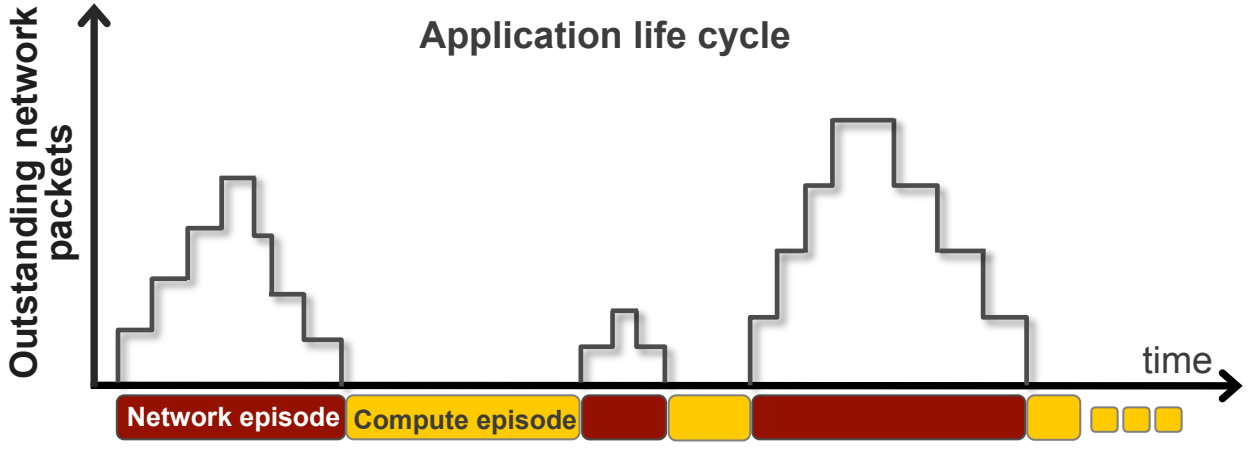
1. Identify LAT/BW sensitive applications
   - Proposes a novel dynamic application classification scheme
2. Design sub-networks based on applications' demand
   - This network architecture is better than a monolithic iso-resource network

Design methodology

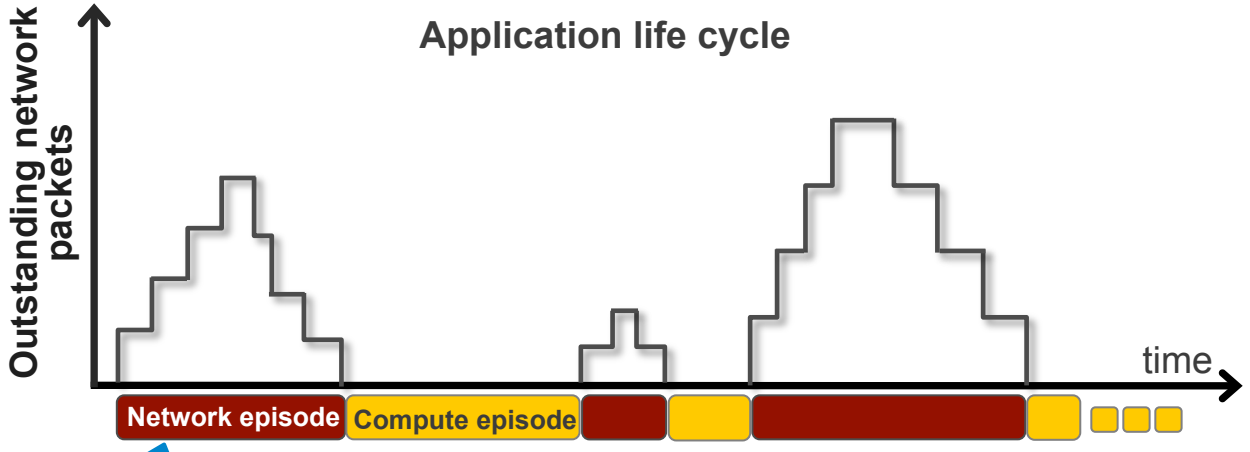


1. Identify LAT/BW sensitive applications
   - Proposes a novel dynamic application classification scheme
2. Design sub-networks based on applications' demand
   - This network architecture is better than a monolithic iso-resource network

Design: Dynamic classification of applications

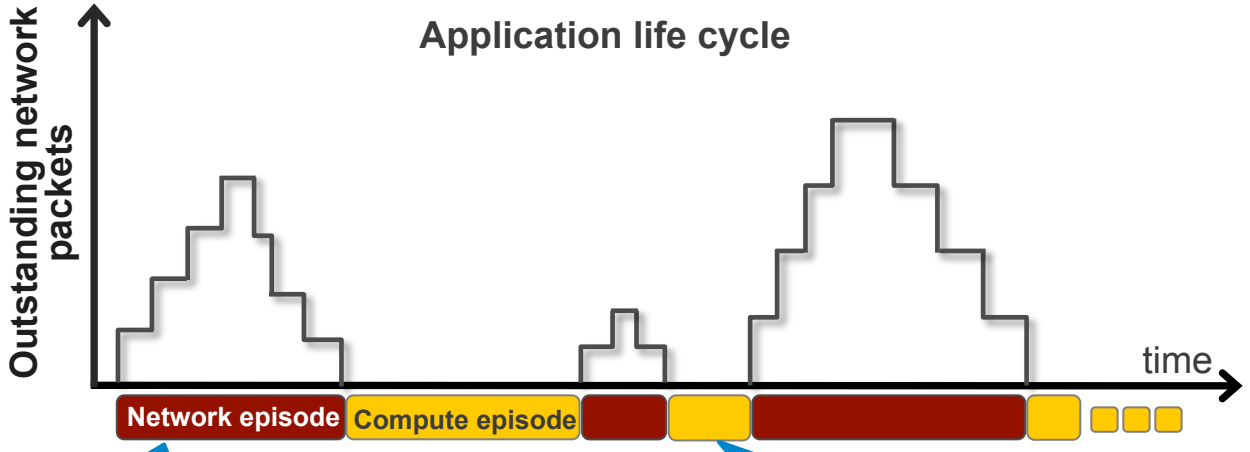


Design: Dynamic classification of applications



- App. has at least one outstanding packet
- Processor is likely stalling → low IPC

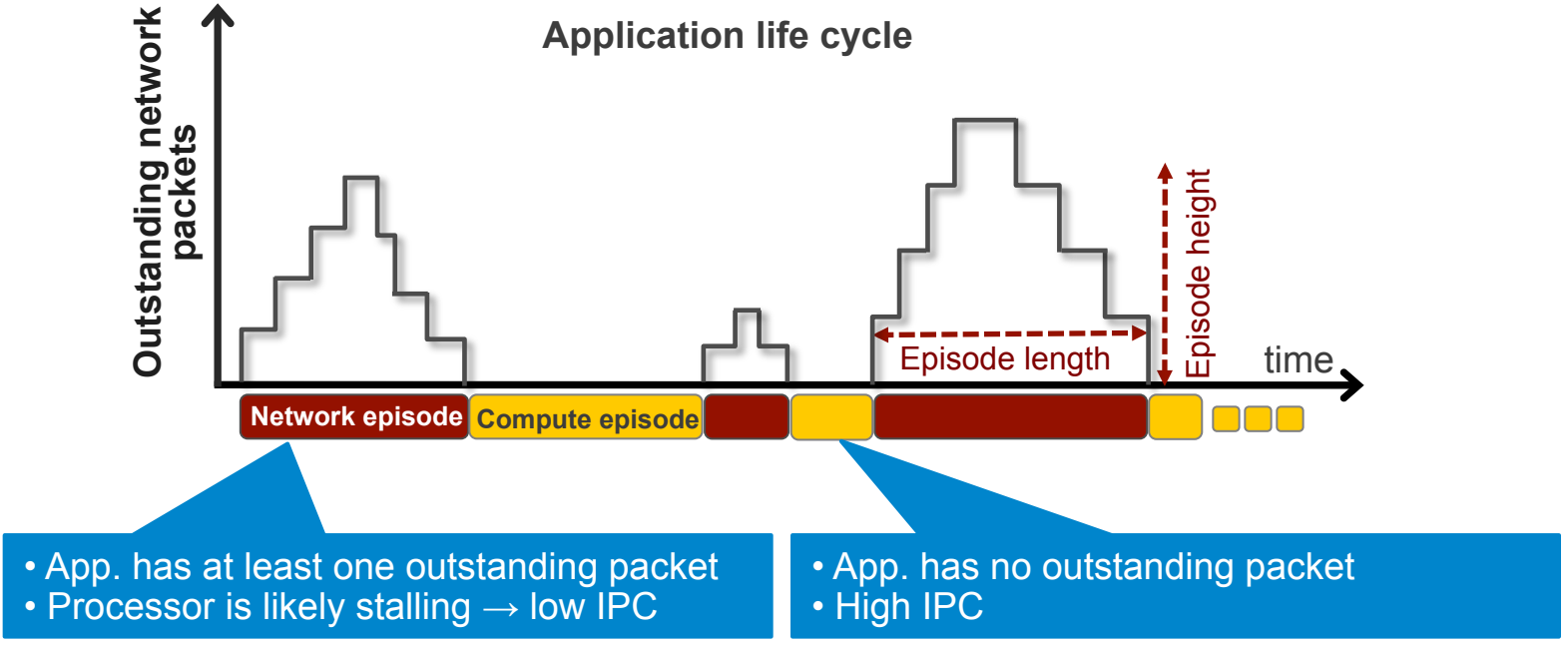
Design: Dynamic classification of applications



- App. has at least one outstanding packet
- Processor is likely stalling → low IPC

- App. has no outstanding packet
- High IPC

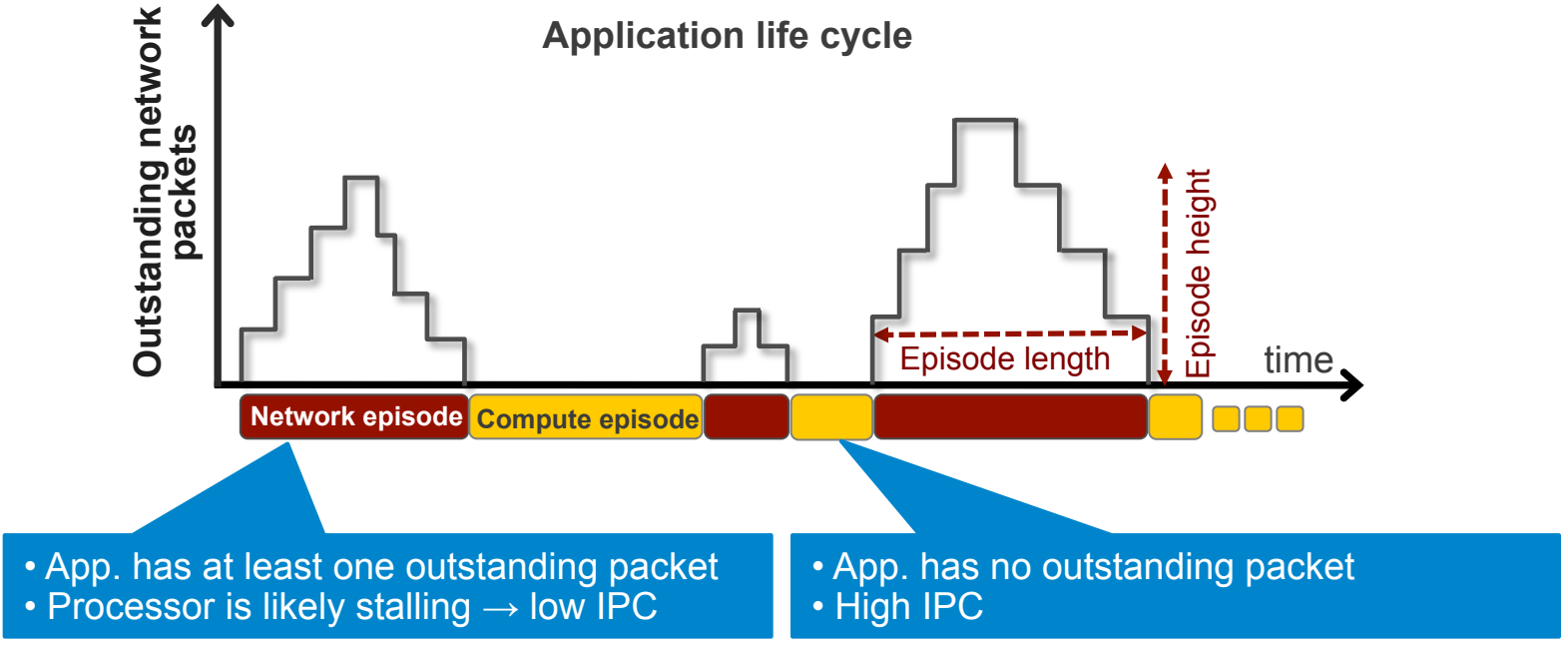
Design: Dynamic classification of applications



Episode length = Number of consecutive cycles there are net. packets

Episode height = Avg. number of L1 packets injected during an episode

Design: Dynamic classification of applications



Short episode ht.: Low MLP, each request is critical (LAT sensitive)
Tall episode ht.: High MLP (BW sensitive)

Short episode len.: Packets are very critical (LAT sensitive)
Long episode len.: Latency tolerant (could be de-prioritized)

Classification and ranking

Classification		Length		
		Long	Medium	Short
Height	Tall	**gems, mcf**	**sphinx, lbm, cactus, xalan**	**sjeng, tonto**
	Medium	**omnetpp, apsi**	**ocean, sjbb, sap, bzip, sjas, soplex, tpc**	applu, perl, barnes, gromacs, namd, calculix, gcc, povray, h264, gobmk, hmmer, astar
	Short	leslie	art, libq, milc, swim	wrf, deal

Classification: LAT/BW

Classification and ranking

Classification

		Length		
		Long	**Medium**	**Short**
Height	**Tall**	**gems, mcf**	**sphinx, lbm, cactus, xalan**	**sjeng, tonto**
	Medium	**omnetpp, apsi**	**ocean, sjbb, sap, bzip, sjas, soplex, tpc**	applu, perl, barnes, gromacs, namd, calculix, gcc, povray, h264, gobmk, hmmer, astar
	Short	leslie	art, libq, milc, swim	wrf, deal

Classification: LAT/BW
Ranking: Sensitivity to LAT/BW

Ranking

		Length		
		Long	**Medium**	**Short**
Height	**High**	**Rank-4**	**Rank-2**	**Rank-1**
	Medium	**Rank-3**	**Rank-2**	Rank-2
	Short	Rank-4	Rank-3	Rank-1

Network design

1N-128

Network design

1N-128

2N-64x256-ST
(Steering)

Network design

1N-128

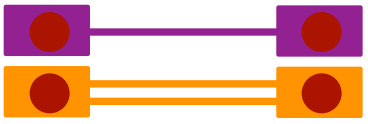
2N-64x256-ST
(Steering)

2N-64x256-ST-RK
(Steering+Ranking)

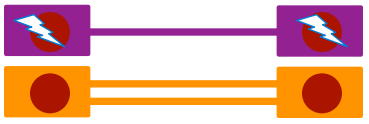
Network design

1N-128

2N-64x256-ST
(Steering)



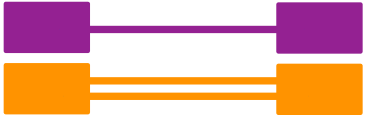
2N-64x256-ST-RK
(Steering+Ranking)



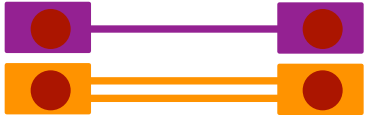
2N-64x256-ST-RK(FS)
(Steering+Ranking and
Frequency Scaling)

Network design

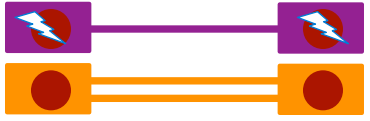
1N-128



2N-64x256-ST
(Steering)

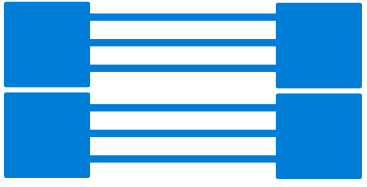


2N-64x256-ST-RK
(Steering+Ranking)



2N-64x256-ST-RK(FS)
(Steering+Ranking and
Frequency Scaling)

1N-256



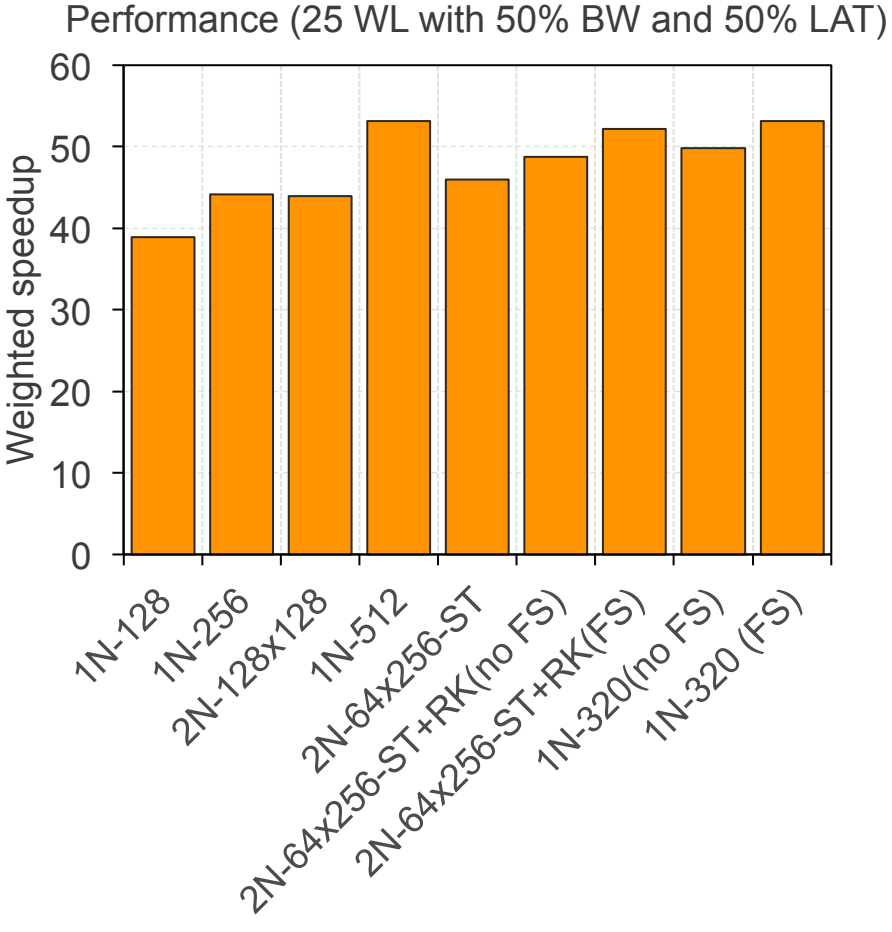
2N-128X128

1N-512
(High BW)

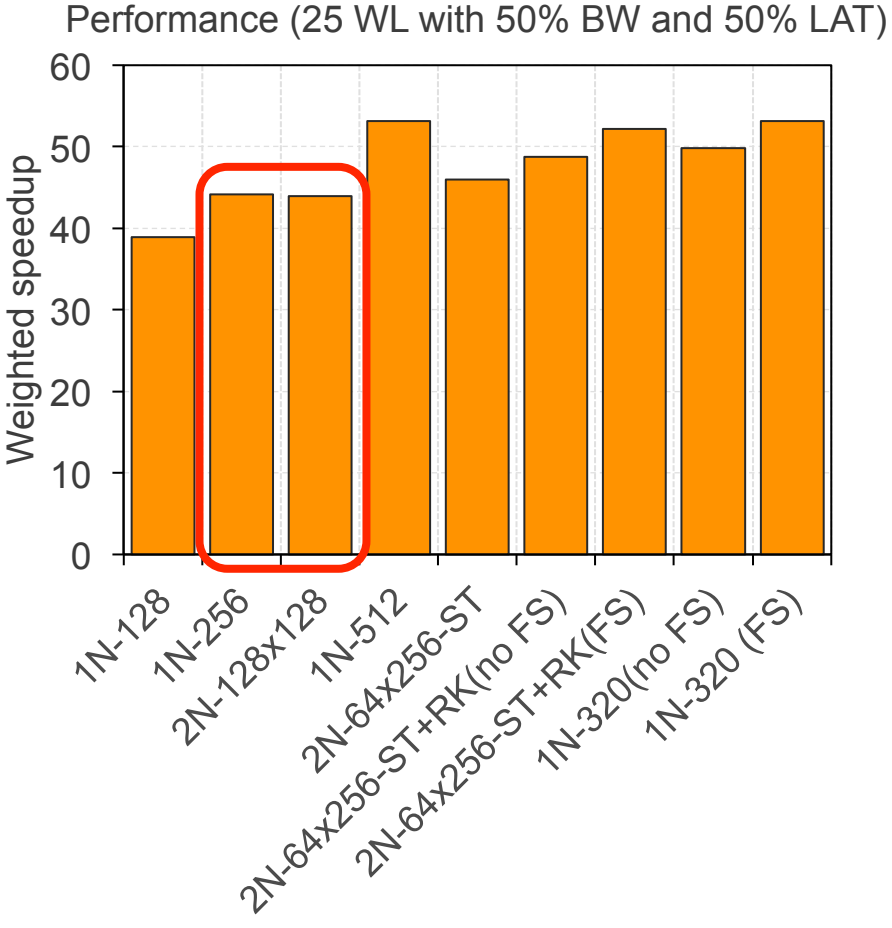
1N-320
(iso-BW)

1N-320(FS)
(iso-resource)

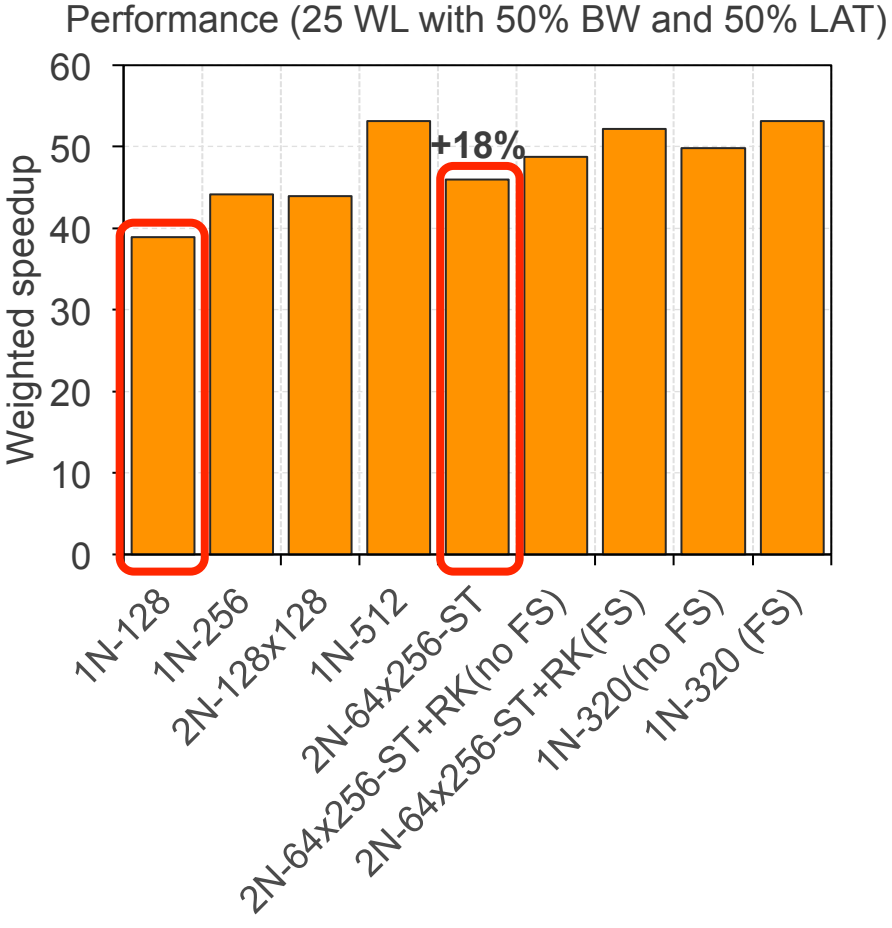
Analysis


Performance (25 WL with 50% BW and 50% LAT)
Weighted speedup
60
50
40
30
20
10
0
1N-128
1N-256
2N-128x128
1N-512
2N-64x256-ST
2N-64x256-ST+RK(no FS)
2N-64x256-ST+RK(FS)
1N-320(no FS)
1N-320 (FS)

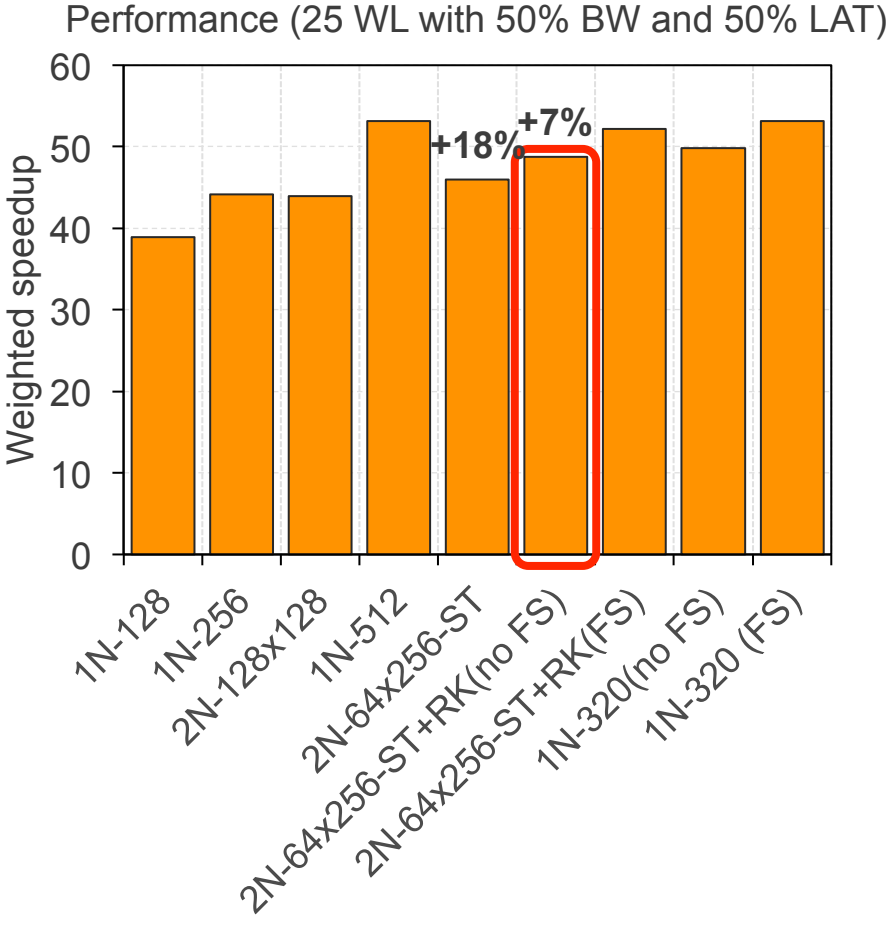
Analysis


Performance (25 WL with 50% BW and 50% LAT)
Weighted speedup
60
50
40
30
20
10
0
1N-128
1N-256
2N-128x128
1N-512
2N-64x256-ST
2N-64x256-ST+RK(no FS)
2N-64x256-ST+RK(FS)
1N-320(no FS)
1N-320 (FS)

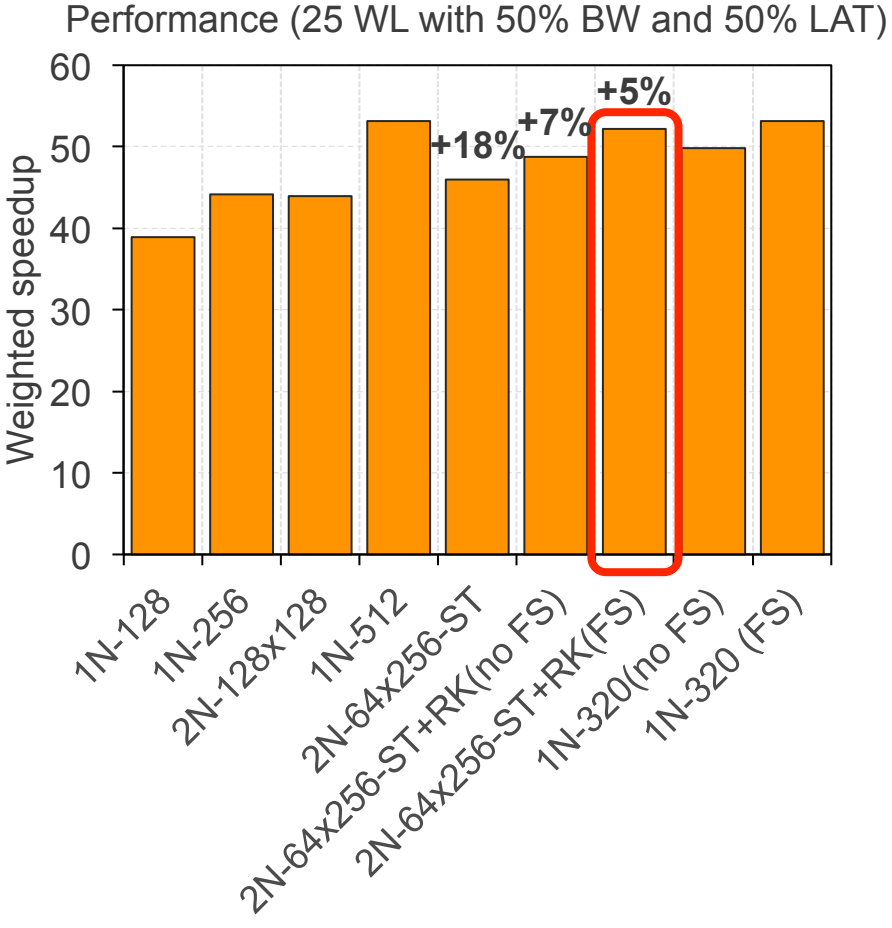
Analysis


Performance (25 WL with 50% BW and 50% LAT)
Weighted speedup
60
50
40
30
20
10
0
+18%
1N-128
1N-256
2N-128x128
1N-512
2N-64x256-ST
2N-64x256-ST+RK(no FS)
2N-64x256-ST+RK(FS)
1N-320(no FS)
1N-320 (FS)

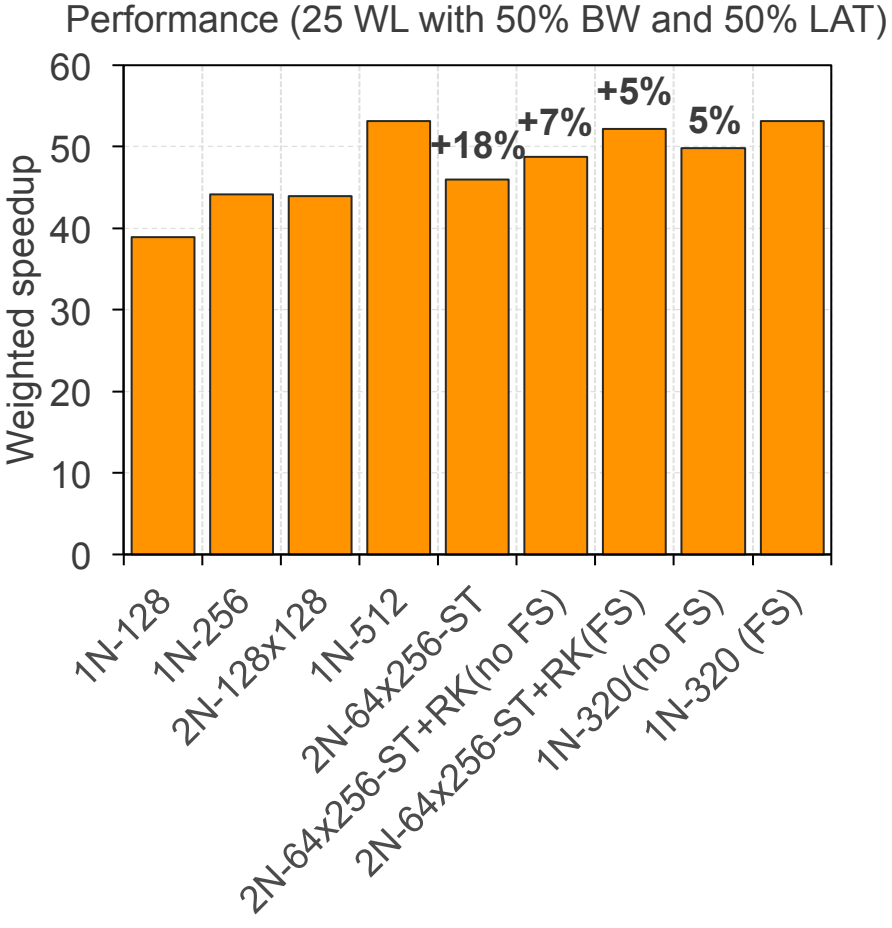
Analysis


Performance (25 WL with 50% BW and 50% LAT)
Weighted speedup
60
50
40
30
20
10
0
+18%
+7%
1N-128
1N-256
2N-128x128
1N-512
2N-64x256-ST
2N-64x256-ST+RK(no FS)
2N-64x256-ST+RK(FS)
1N-320(no FS)
1N-320 (FS)

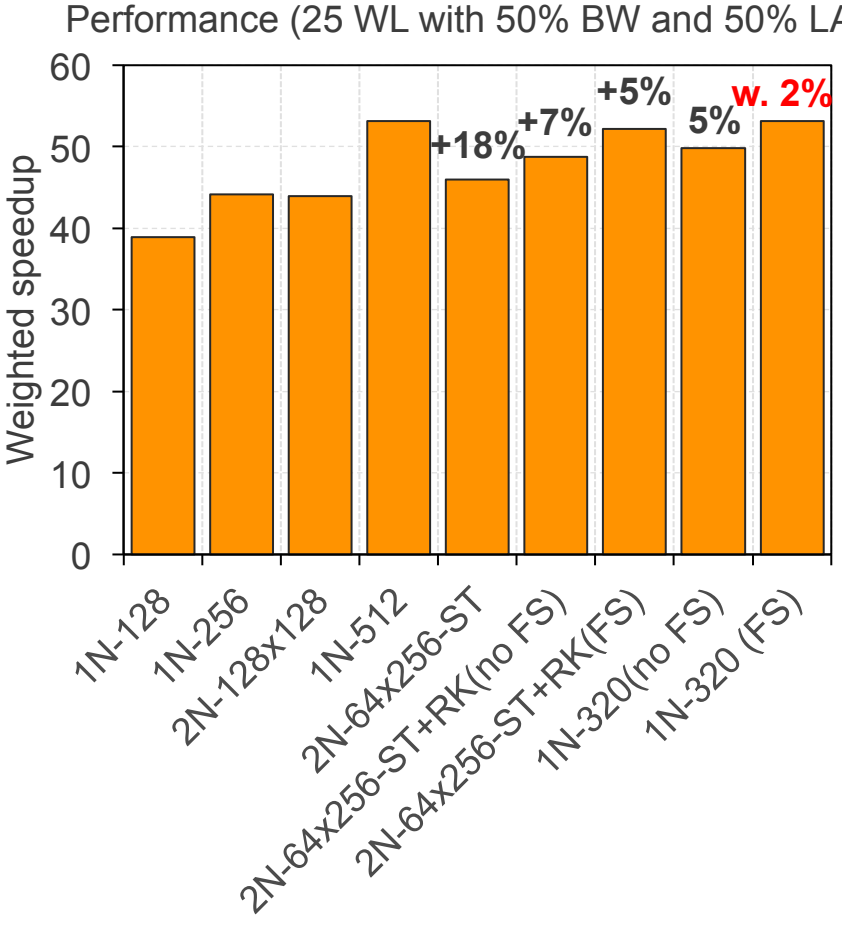
Analysis


Performance (25 WL with 50% BW and 50% LAT)
Weighted speedup
60
50
40
30
20
10
0
+18%
+7%
+5%
1N-128
1N-256
2N-128x128
1N-512
2N-64x256-ST
2N-64x256-ST+RK(no FS)
2N-64x256-ST+RK(FS)
1N-320(no FS)
1N-320 (FS)

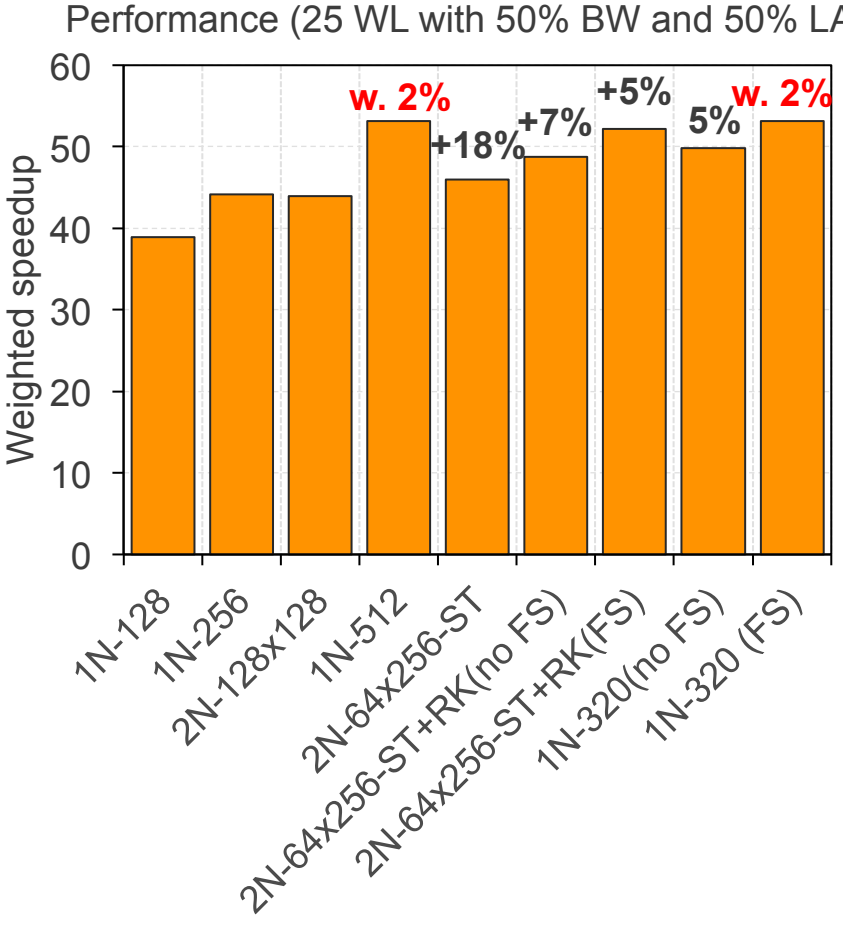
Analysis


Performance (25 WL with 50% BW and 50% LAT)
Weighted speedup
60
50
40
30
20
10
0
+18%
+7%
+5%
5%
1N-128
1N-256
2N-128x128
1N-512
2N-64x256-ST
2N-64x256-ST+RK(no FS)
2N-64x256-ST+RK(FS)
1N-320(no FS)
1N-320 (FS)

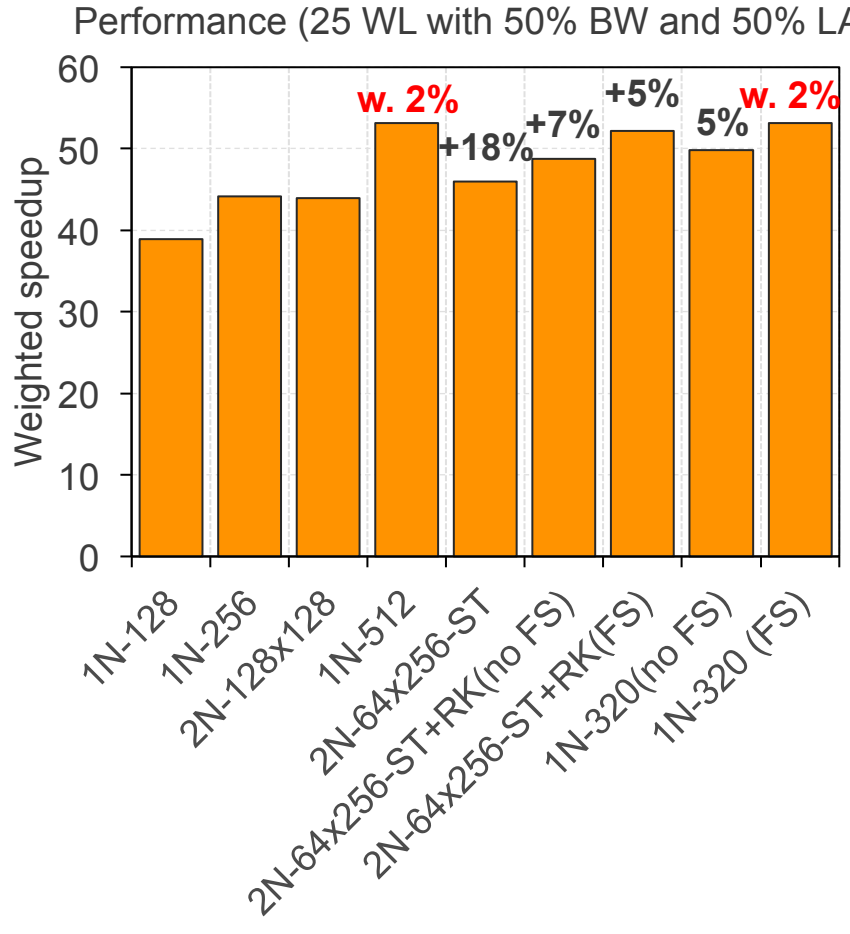
Analysis

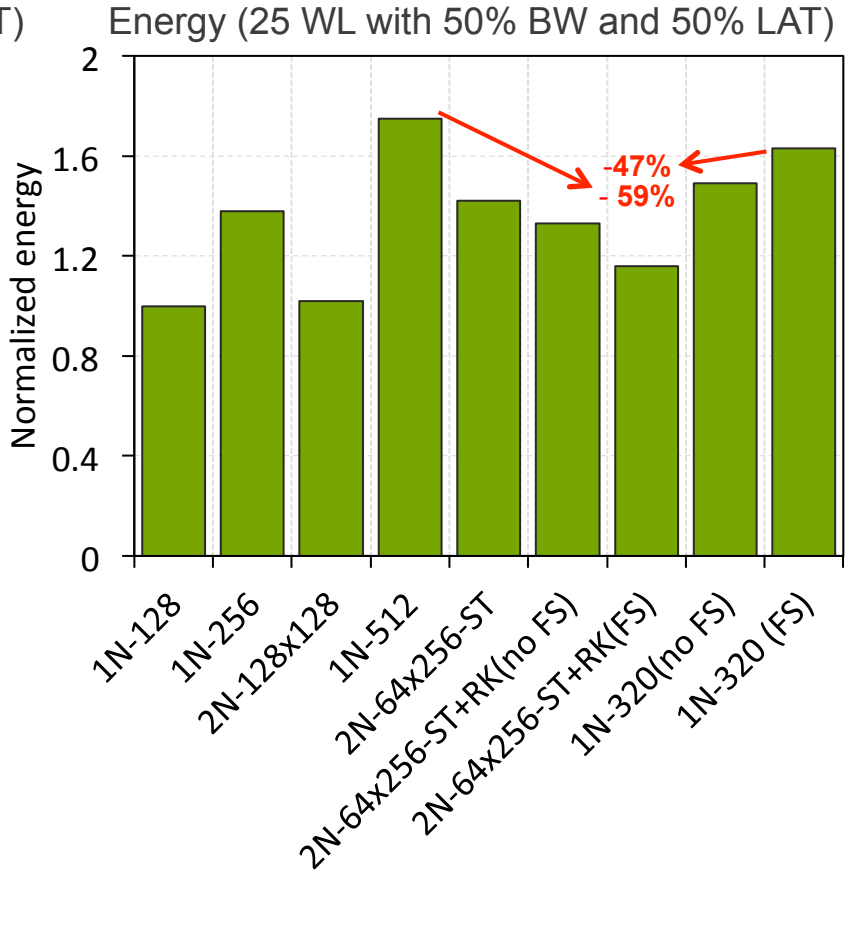

Performance (25 WL with 50% BW and 50% LAT)
Weighted speedup
60
50
40
30
20
10
0
+18%
+7%
+5%
5%
w. 2%
1N-128
1N-256
2N-128x128
1N-512
2N-64x256-ST
2N-64x256-ST+RK(no FS)
2N-64x256-ST+RK(FS)
1N-320(no FS)
1N-320 (FS)

Analysis

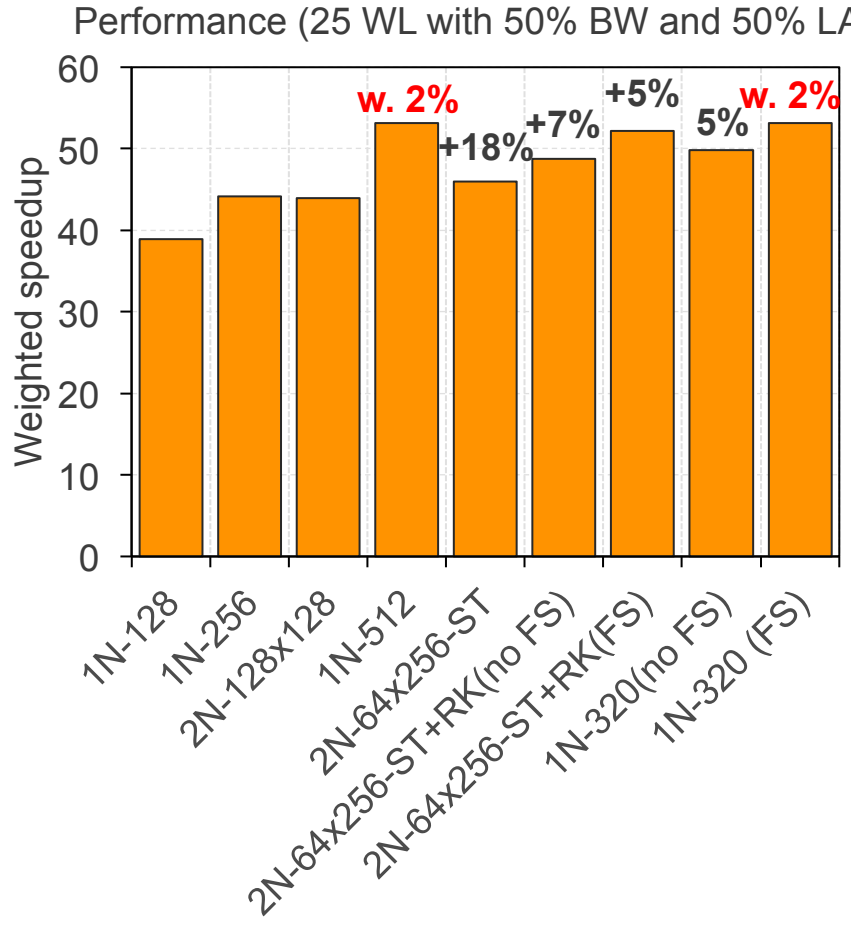

Performance (25 WL with 50% BW and 50% LAT)
Weighted speedup
60
50
40
30
20
10
0
w. 2%
+18%
+7%
+5%
5%
w. 2%
1N-128
1N-256
2N-128x128
1N-512
2N-64x256-ST
2N-64x256-ST+RK(no FS)
2N-64x256-ST+RK(FS)
1N-320(no FS)
1N-320 (FS)

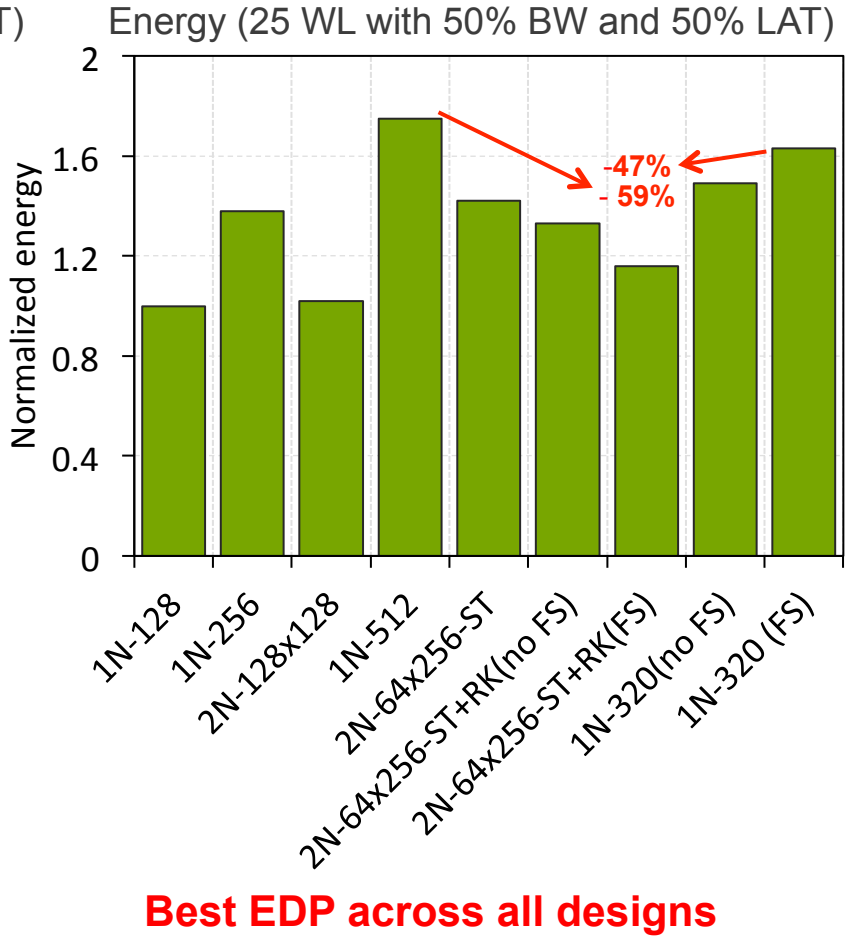
Analysis


Performance (25 WL with 50% BW and 50% LAT)
Weighted speedup
0
10
20
30
40
50
60
w. 2%
+18%
+7%
+5%
5%
w. 2%
1N-128
1N-256
2N-128x128
1N-512
2N-64x256-ST
2N-64x256-ST+RK(no FS)
2N-64x256-ST+RK(FS)
1N-320(no FS)
1N-320 (FS)


Energy (25 WL with 50% BW and 50% LAT)
Normalized energy
0
0.4
0.8
1.2
1.6
2
-47%
- 59%
1N-128
1N-256
2N-128x128
1N-512
2N-64x256-ST
2N-64x256-ST+RK(no FS)
2N-64x256-ST+RK(FS)
1N-320(no FS)
1N-320 (FS)

Analysis





Best EDP across all designs

Conclusions

• **Problem**: Current day NoC designs are **agnostic** to application requirements and are provisioned for the general case or worst case. Applications have widely differing demands from the network

• **Our goal**: To design a NoC that can satisfy the **diverse** dynamic performance requirements of applications

• **Observation**: Applications can be divided into two general classes in terms of their requirements from the network: **bandwidth-sensitive** and **latency-sensitive**
  - Not all applications are equally sensitive to bandwidth and latency

• **Key idea**: Design two NoC
  - Each sub-network customized for either BW or LAT sensitive applications
  - Propose **metrics** to classify applications as BW or LAT sensitive
  - Prioritize applications' packets within the sub-networks based on their sensitivity

• **Network design**: BW optimized network has wider link width but operates at a lower frequency and LAT optimized network has narrow link width but operates at a higher frequency

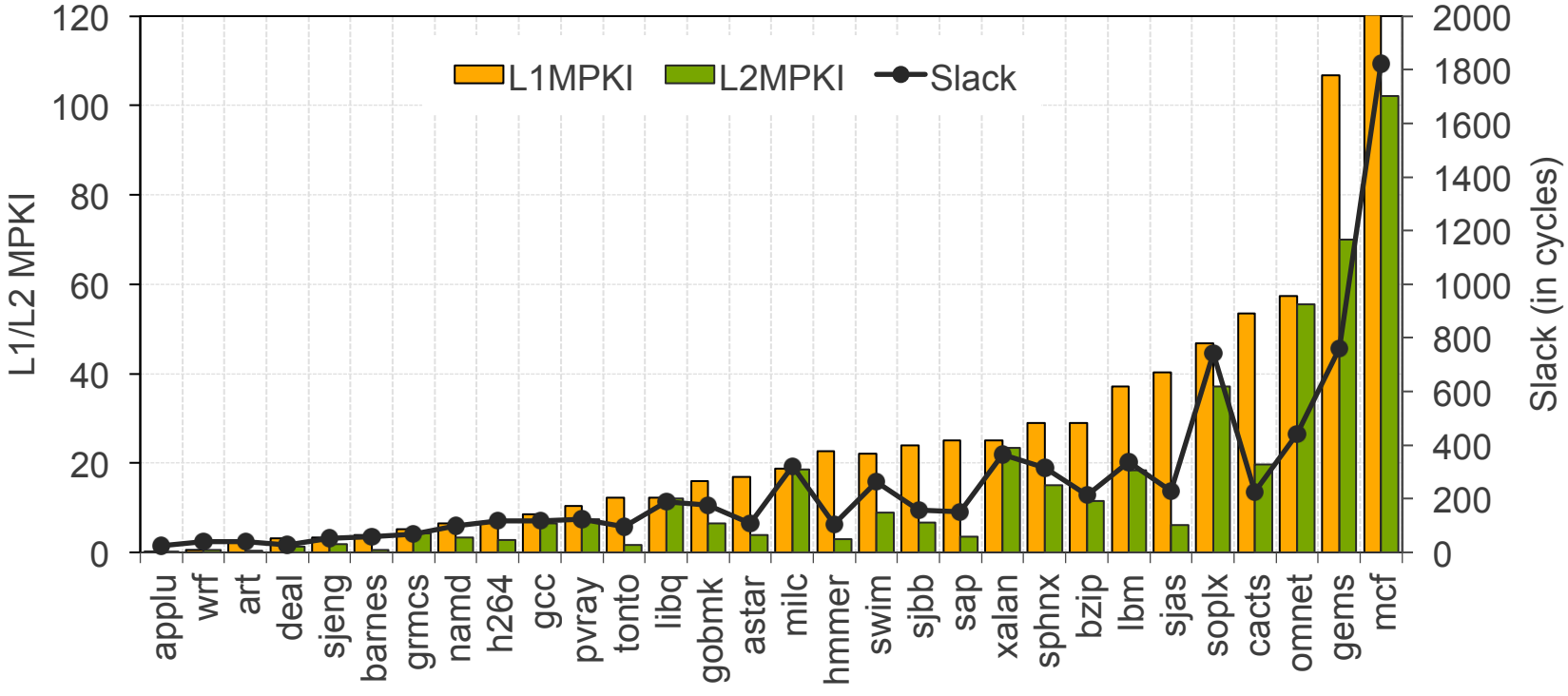
• **Results**: Our proposal is significantly better when compared to an iso-resource monolithic network (5%/3% weighted/instruction throughput improvement and 31% energy reduction)

Thank you

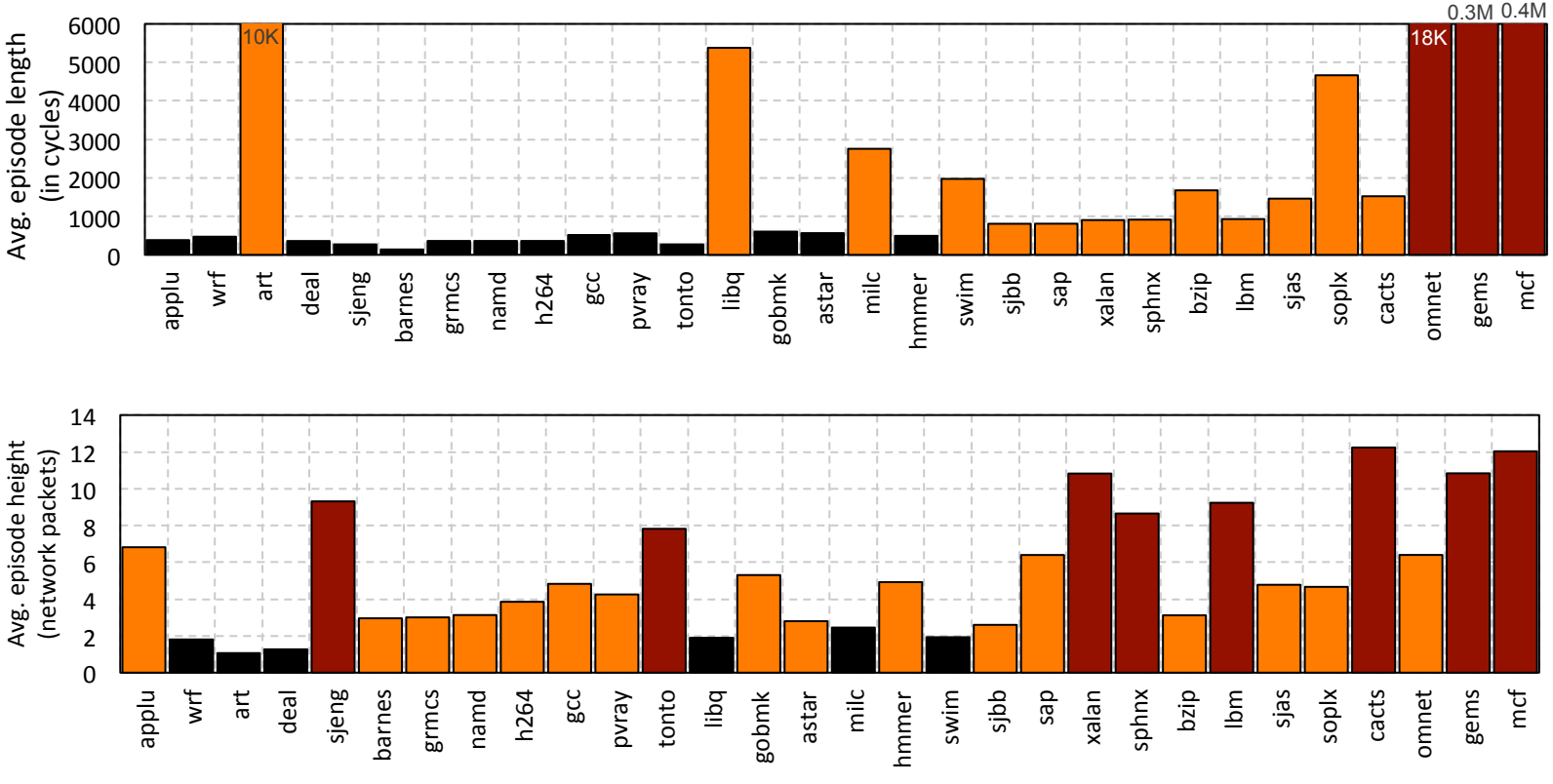
Asit Mishra
asit.k.mishra@intel.com

Backup Slides . . .

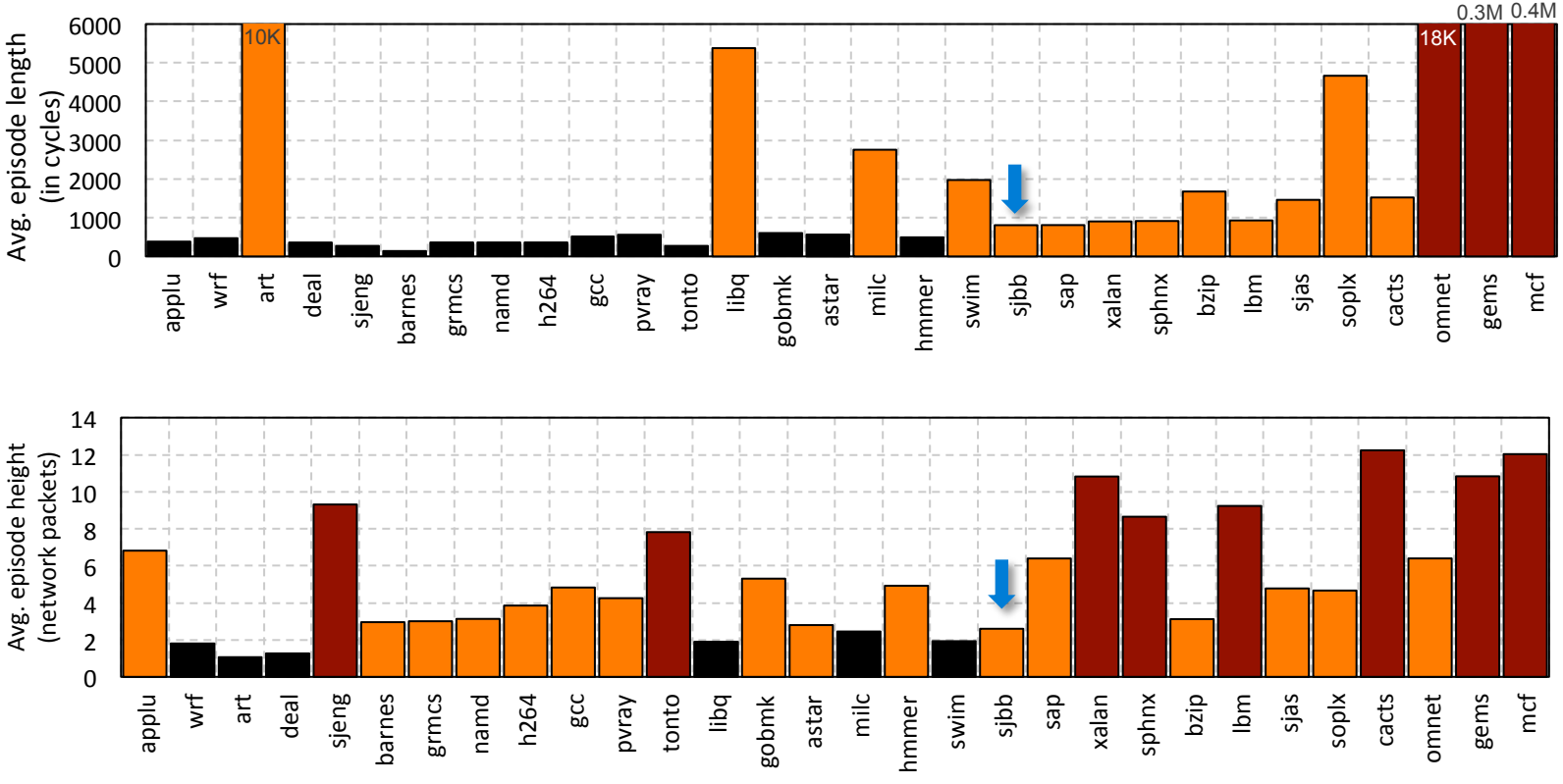
Other metrics considered for application classification



Analysis of network episode length and height



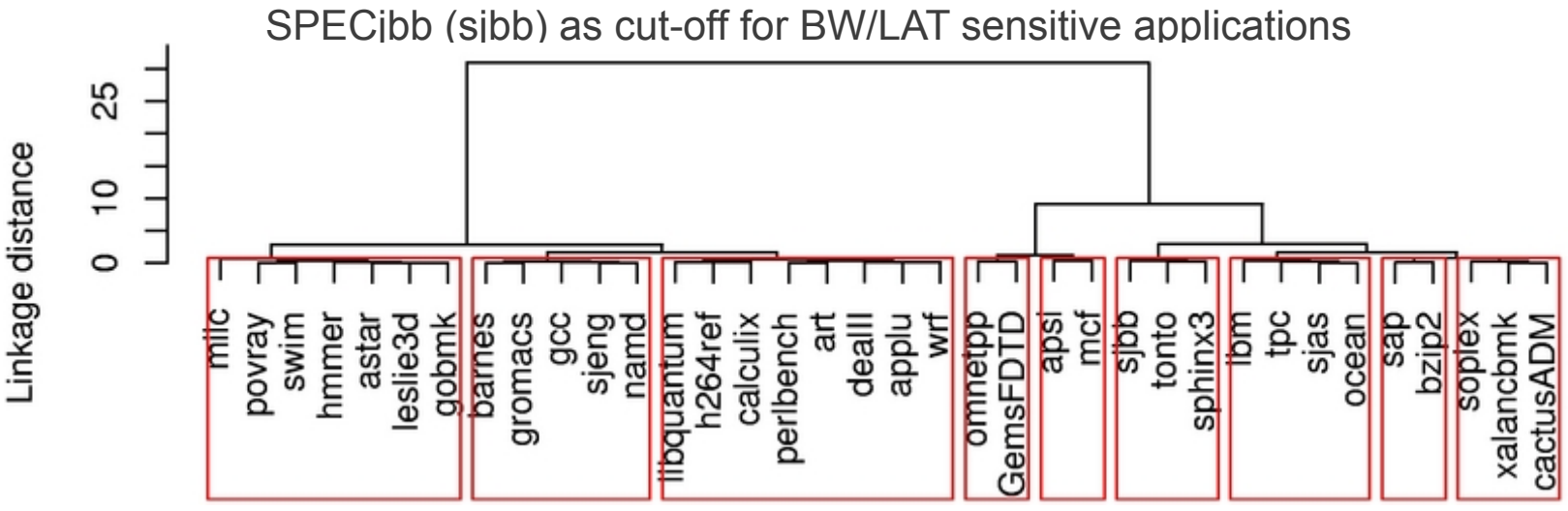
Analysis of network episode length and height



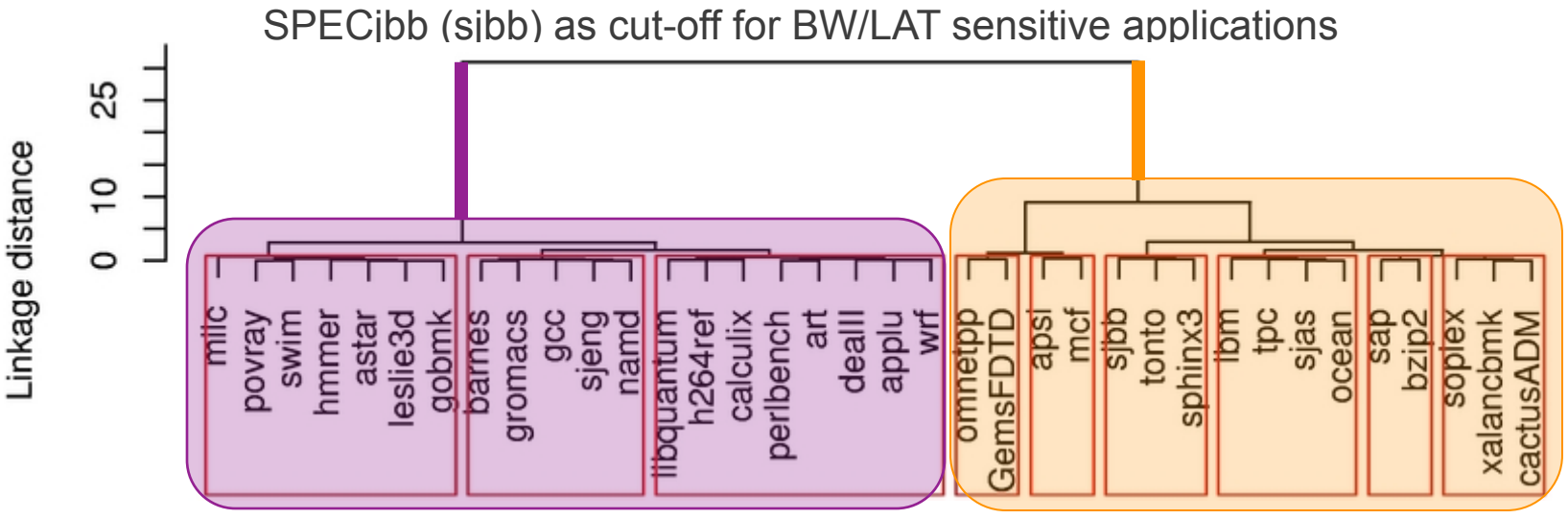
Empirical results to support the classification

SPECjbb (sjbb) as cut-off for BW/LAT sensitive applications

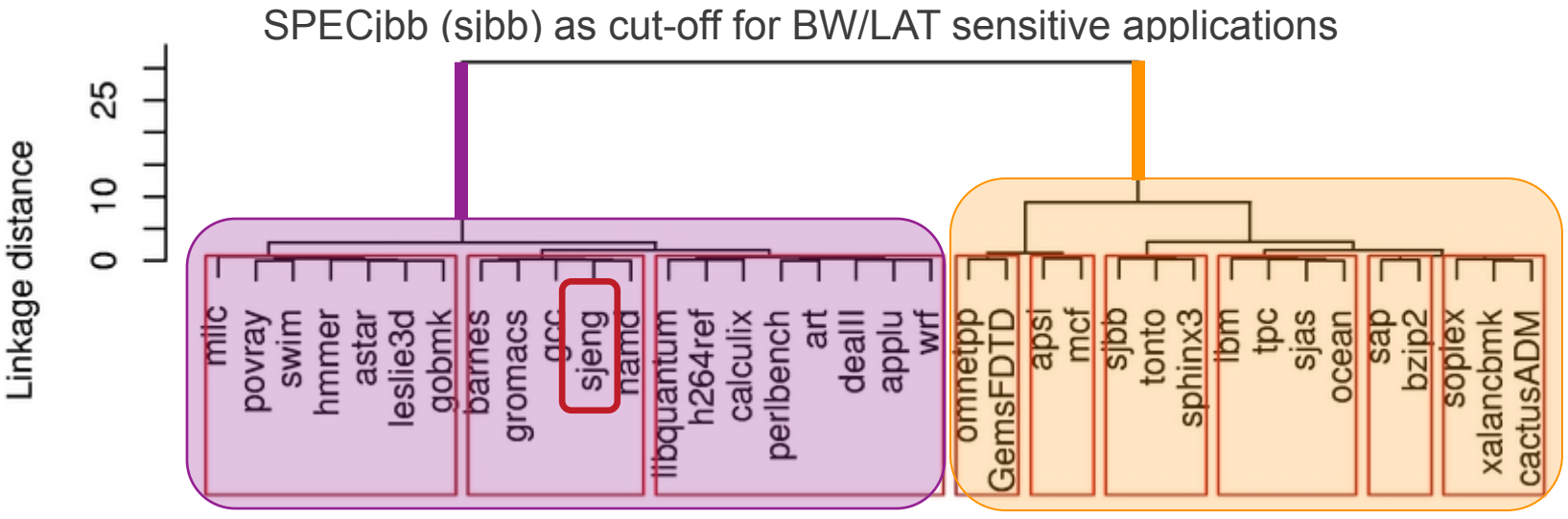
Empirical results to support the classification



Empirical results to support the classification

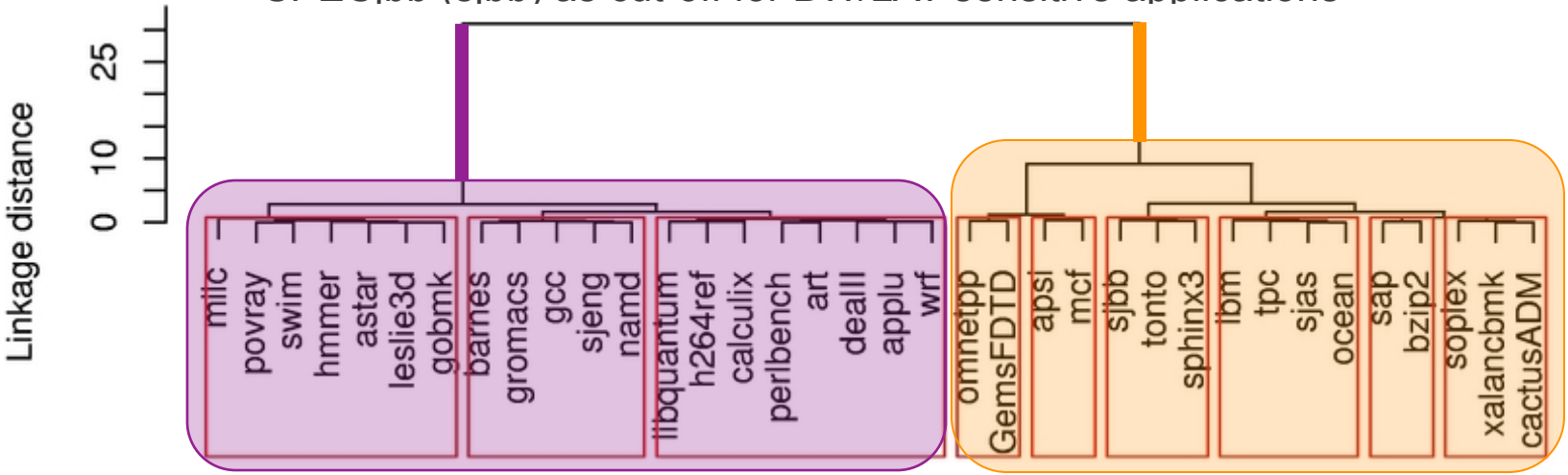


Empirical results to support the classification

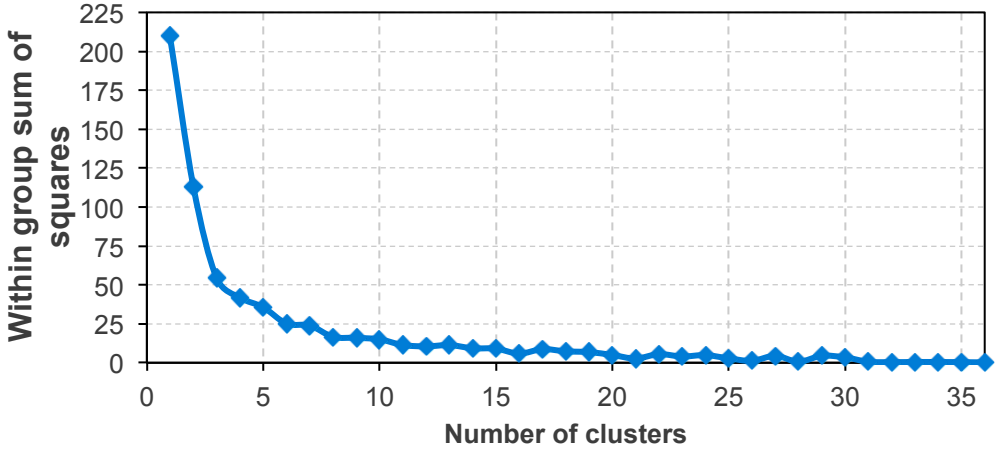


Empirical results to support the classification

SPECjbb (sjbb) as cut-off for BW/LAT sensitive applications

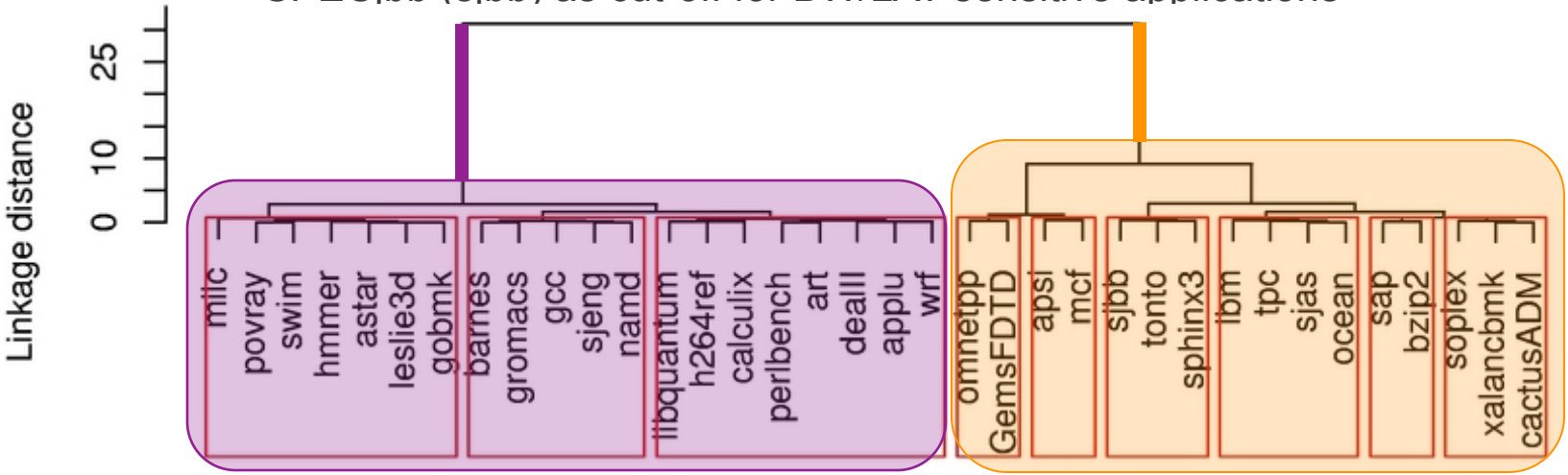


Why 9 clusters?

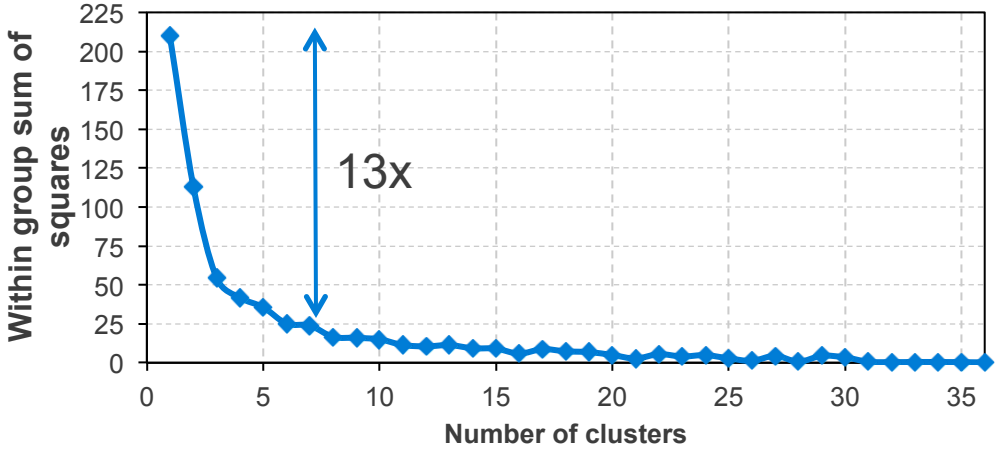


Empirical results to support the classification

SPECjbb (sjbb) as cut-off for BW/LAT sensitive applications

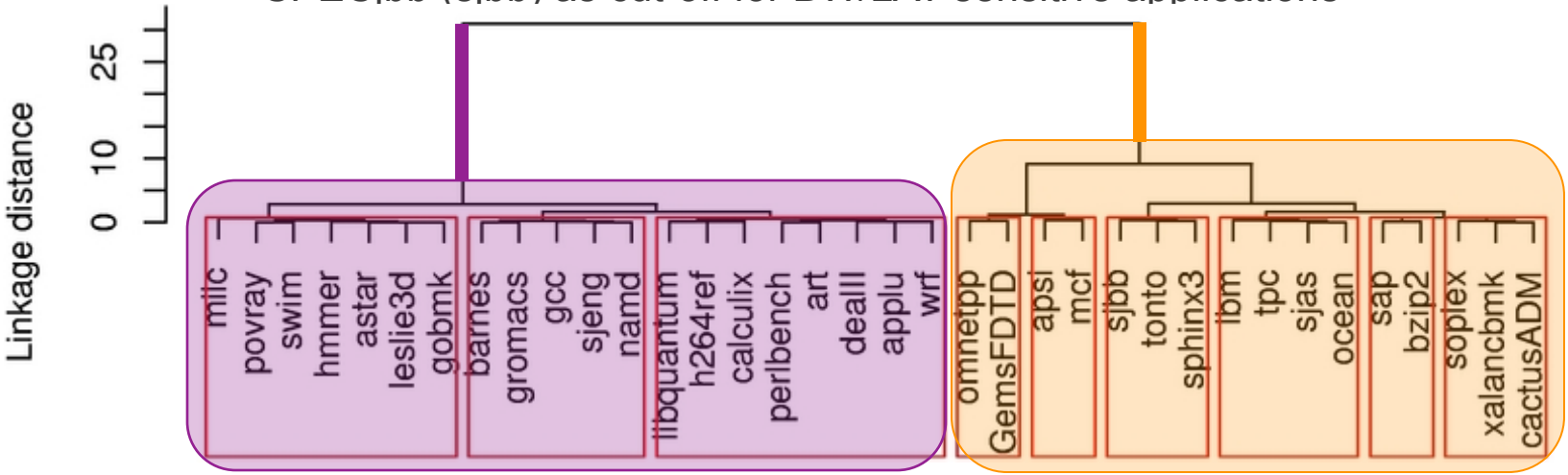


Why 9 clusters?

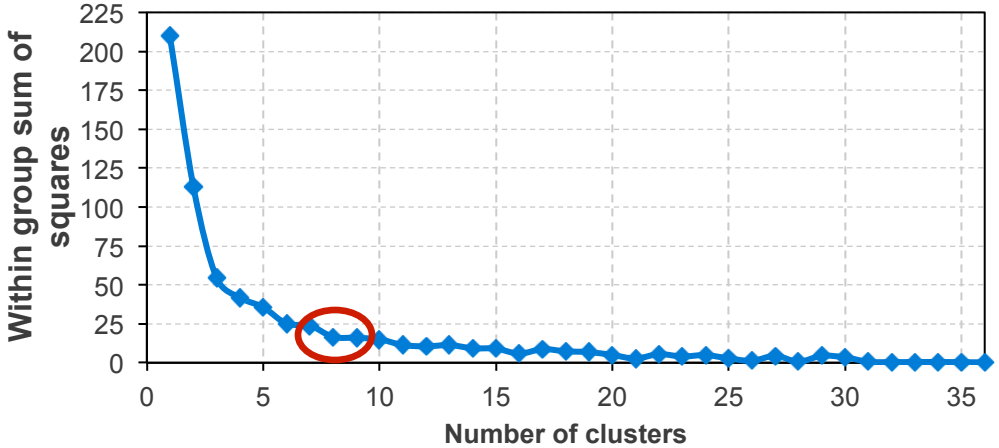


Empirical results to support the classification

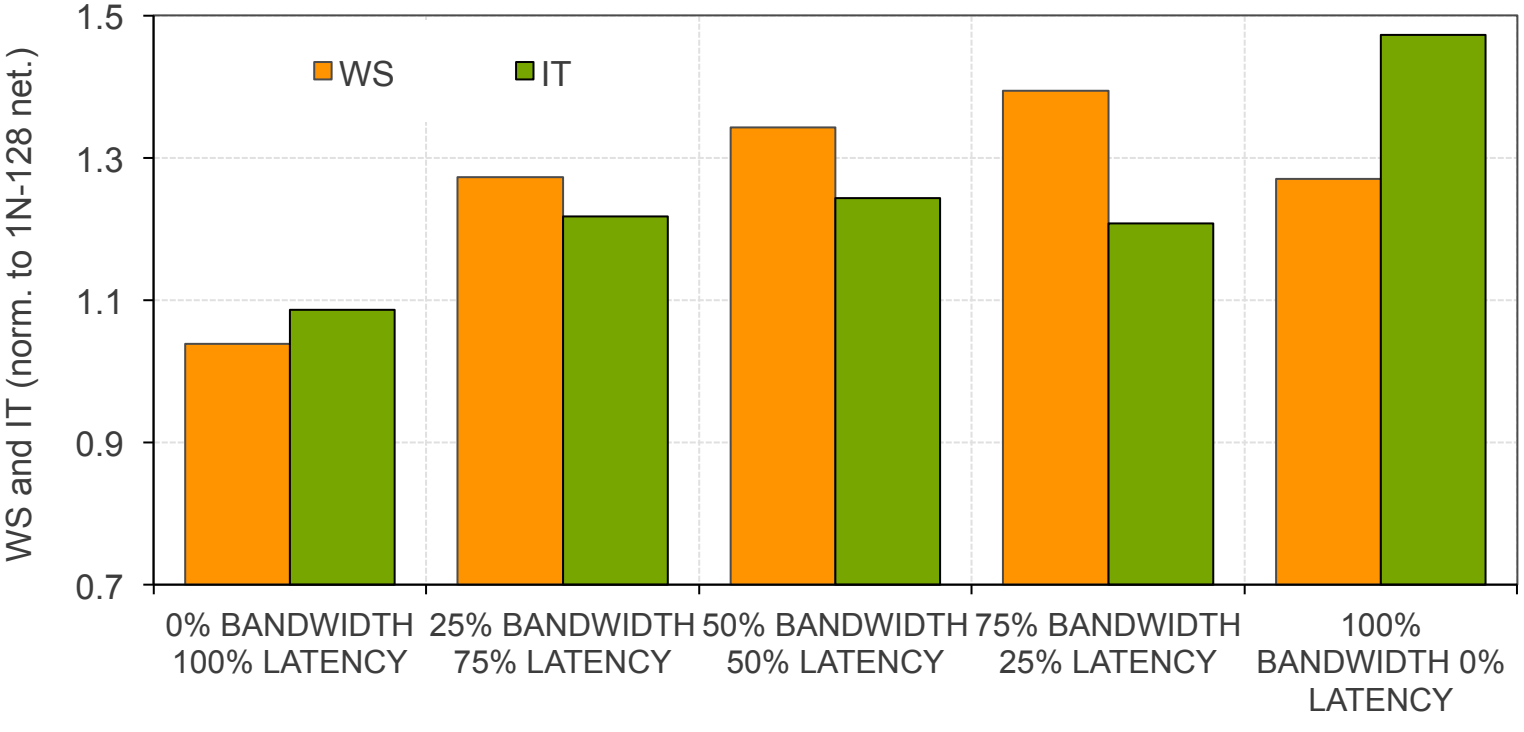
SPECjbb (sjbb) as cut-off for BW/LAT sensitive applications



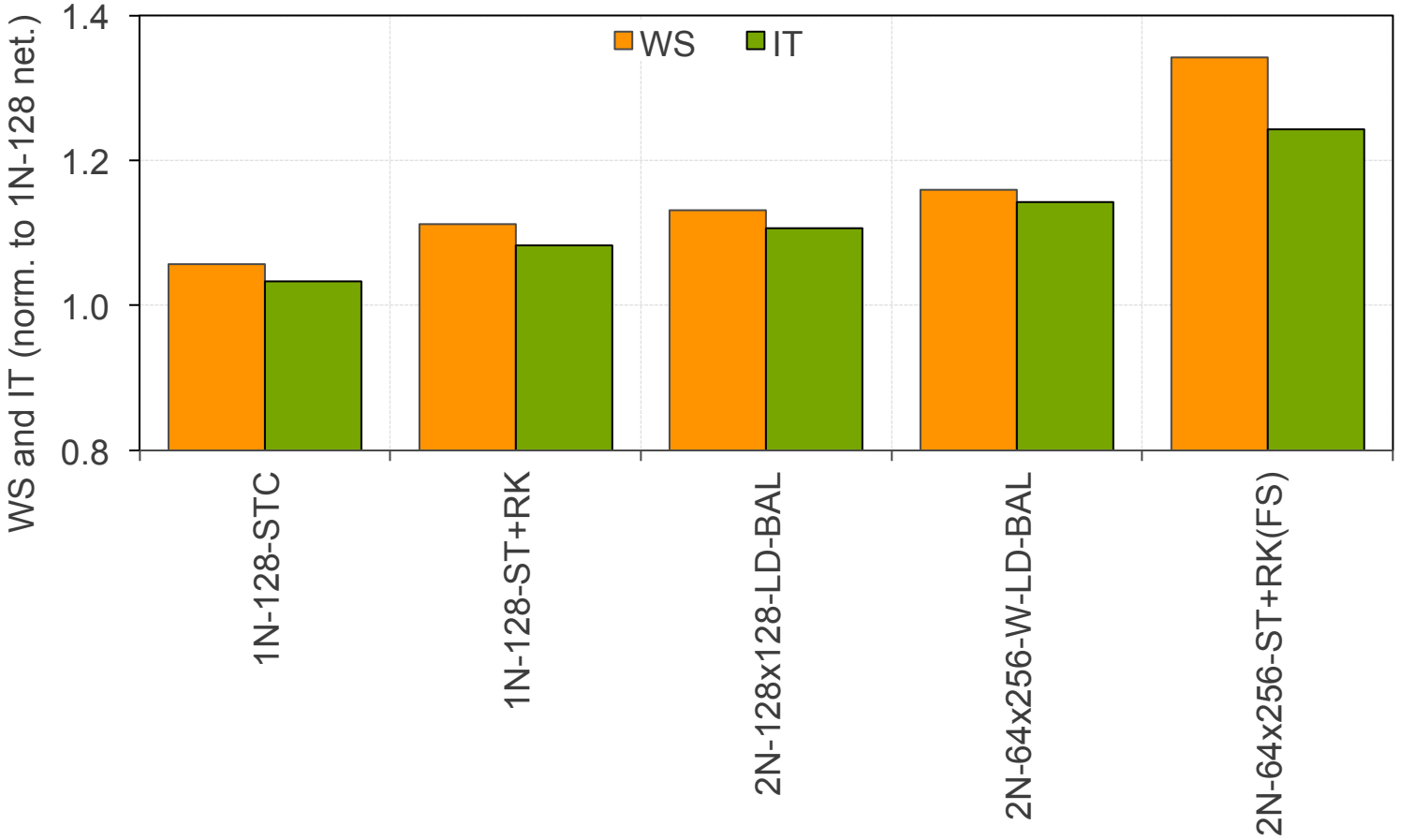
Why 9 clusters?



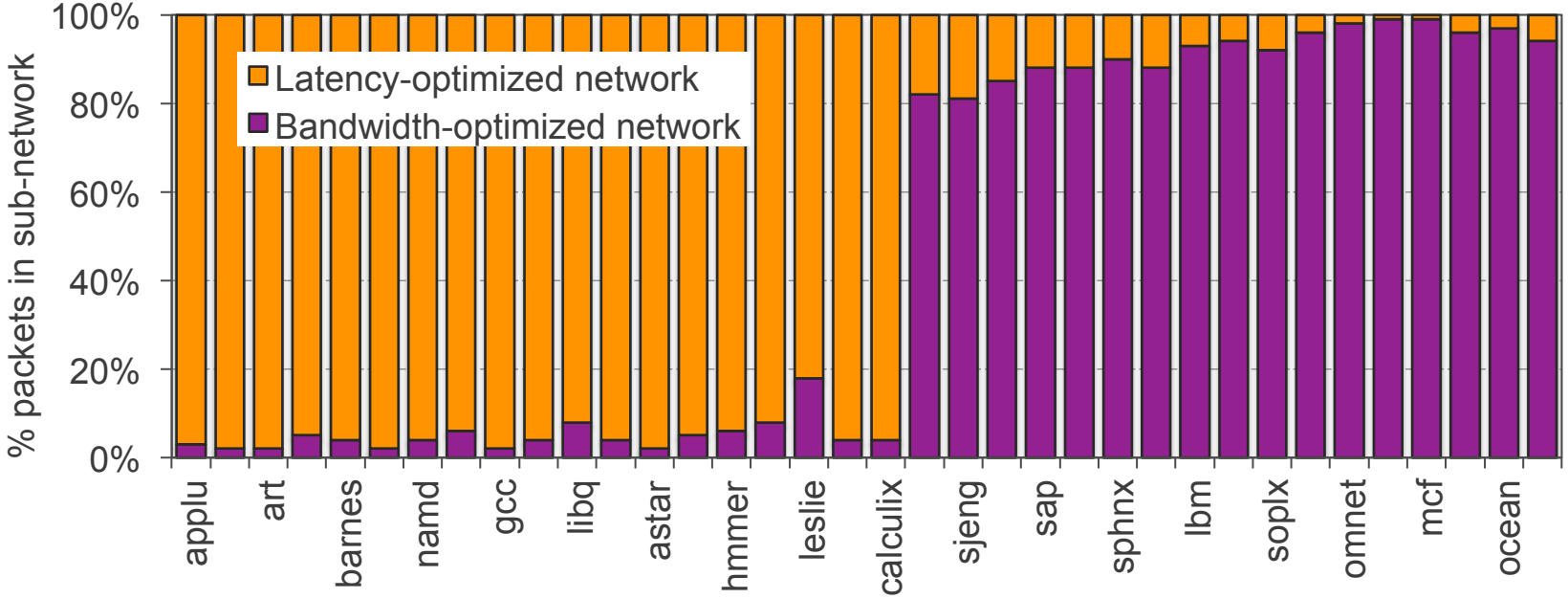
Analysis with varying workload combinations



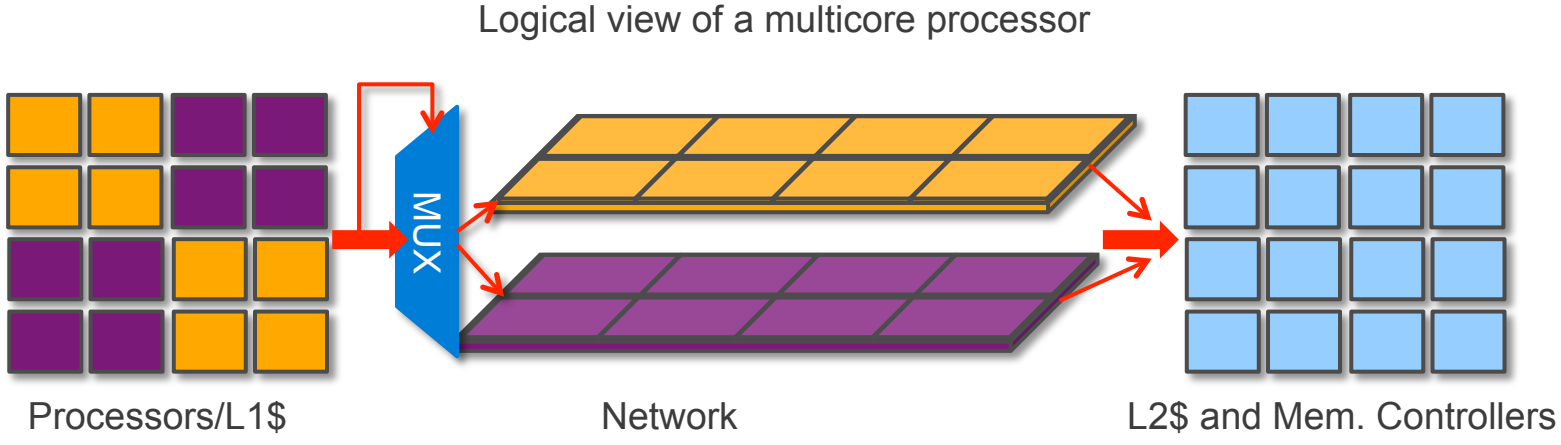
Comparison to prior works



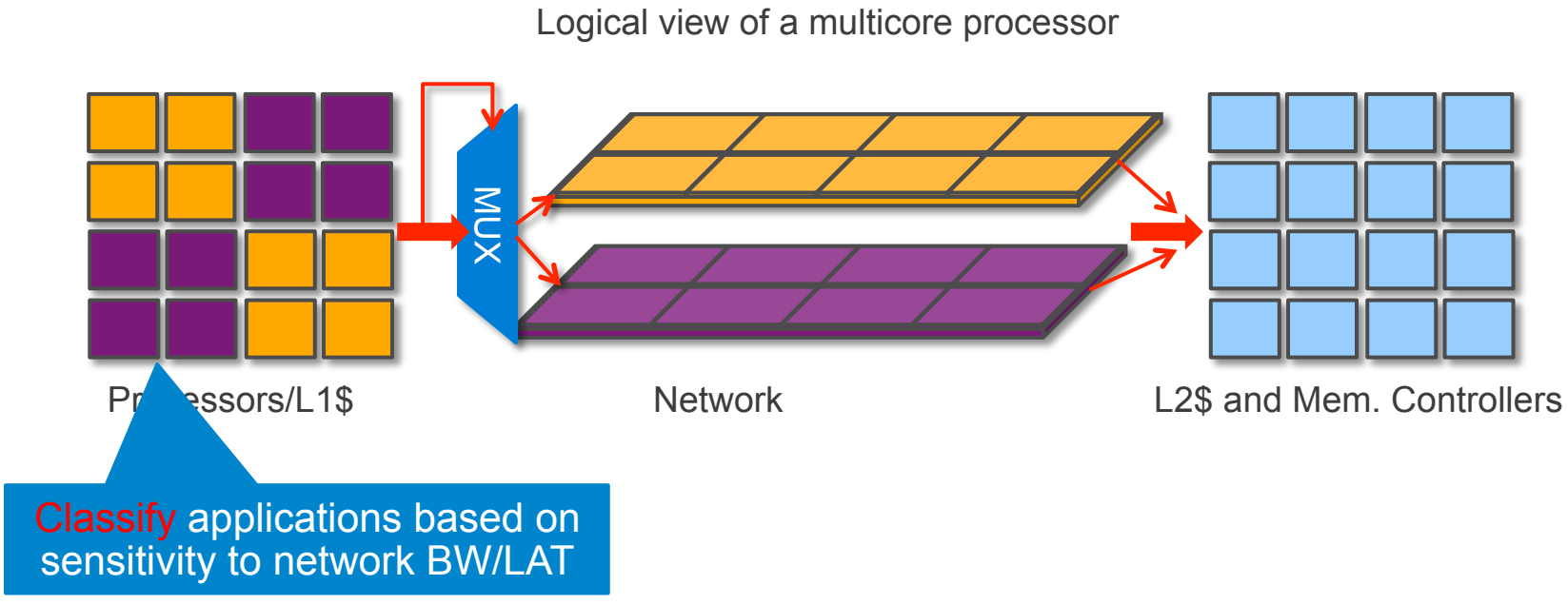
Dynamic steering of packets



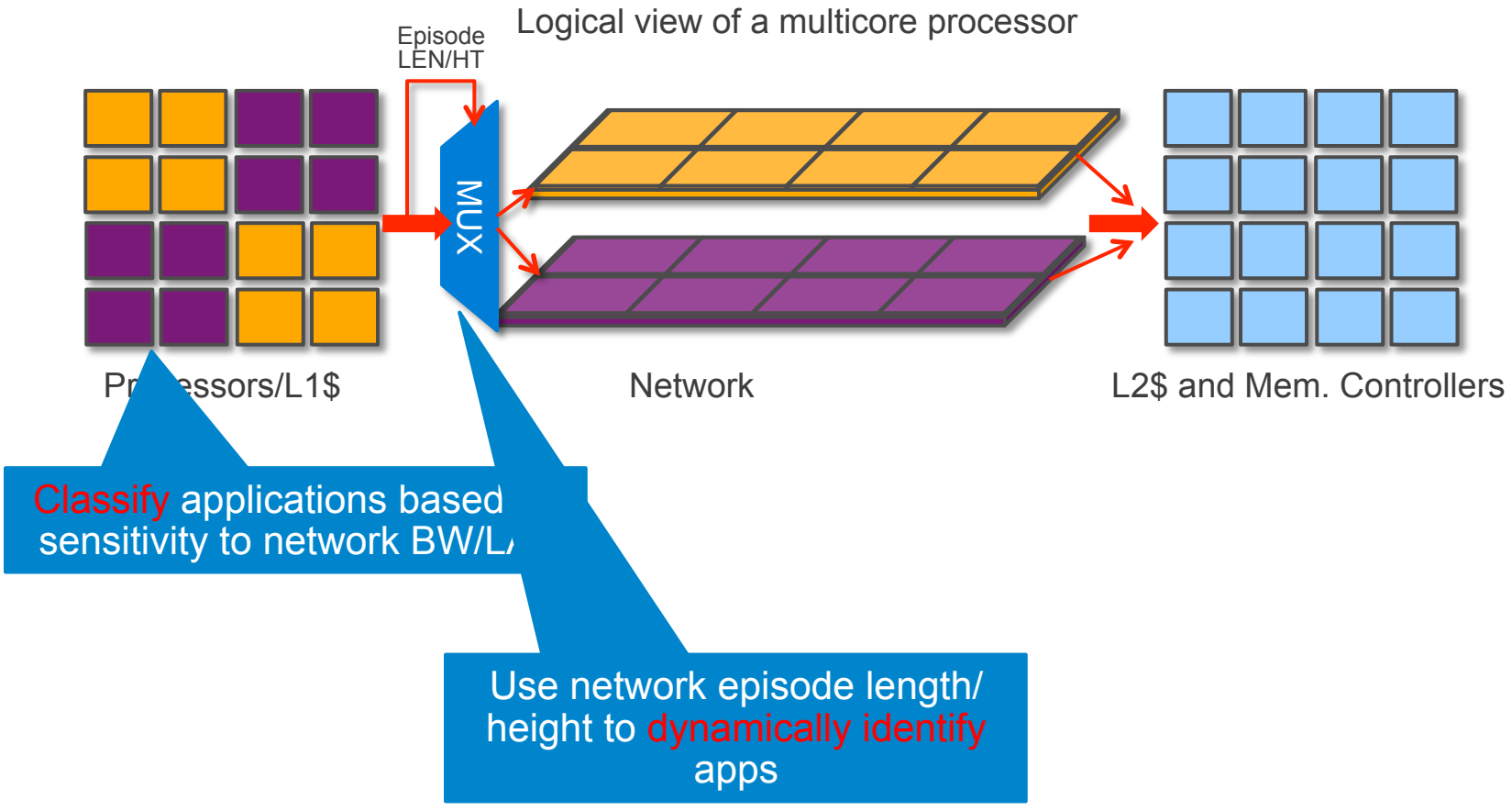
Design: Putting it all together



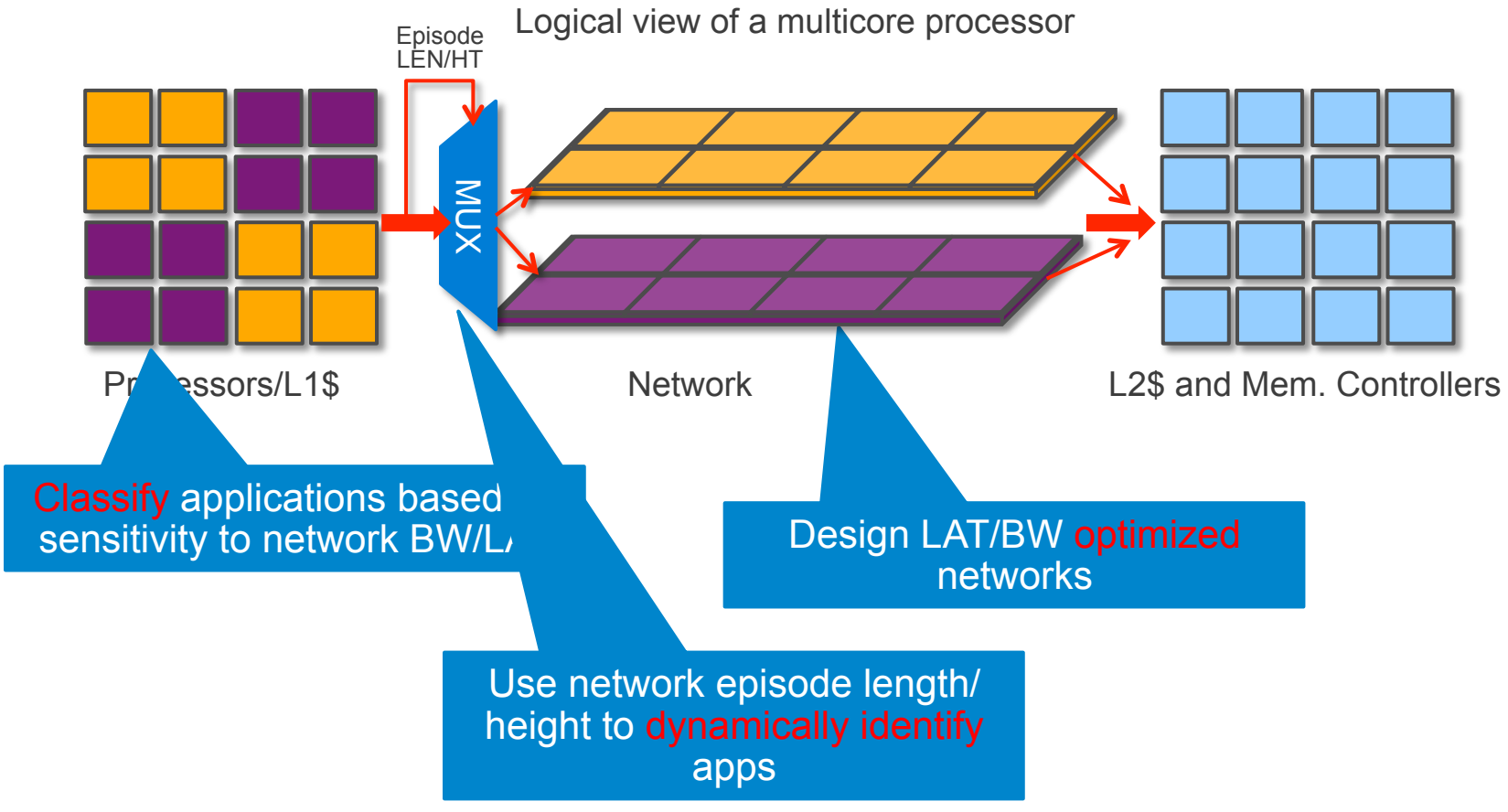
Design: Putting it all together



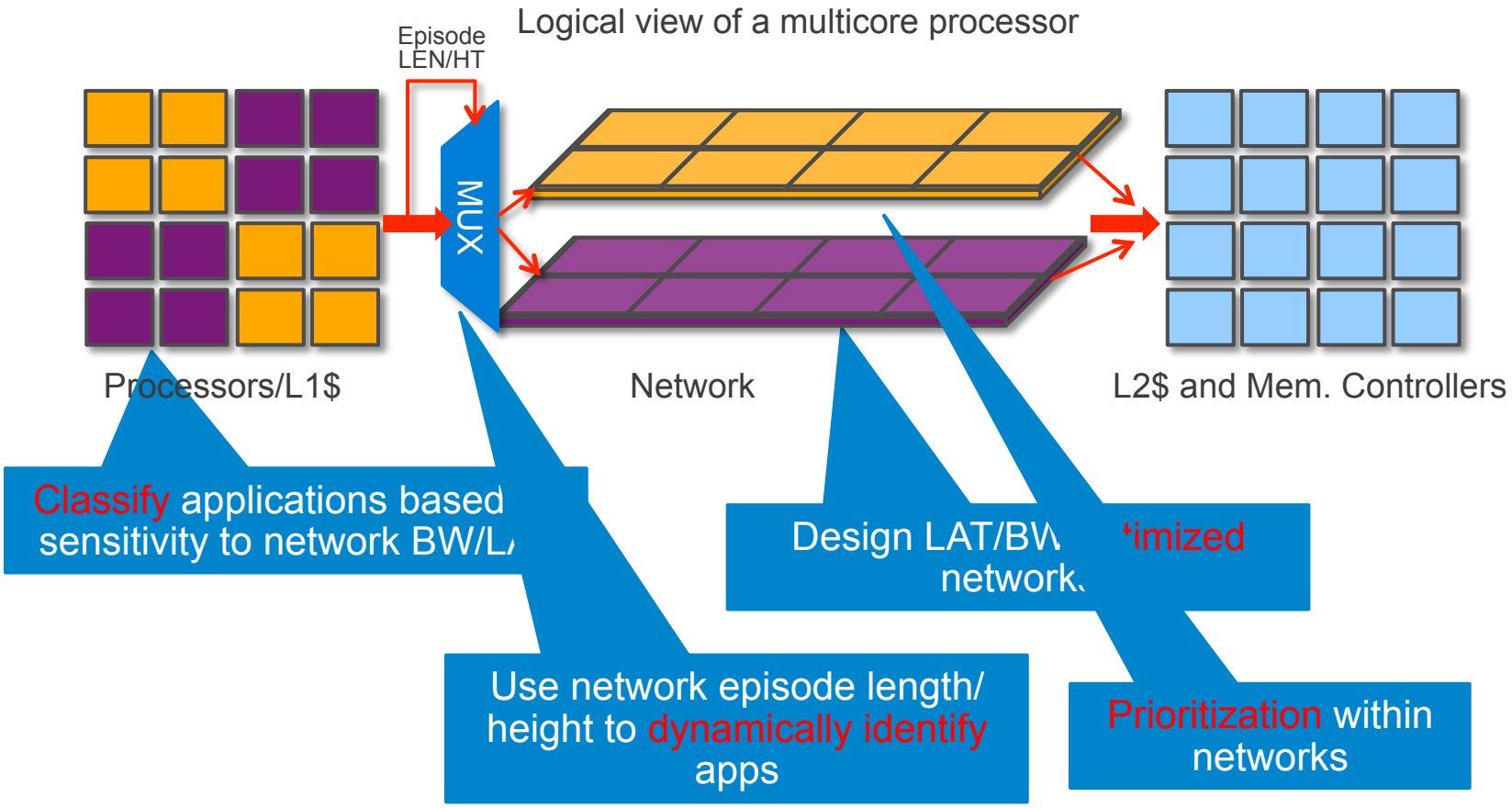
Design: Putting it all together



Design: Putting it all together



Design: Putting it all together



Summary

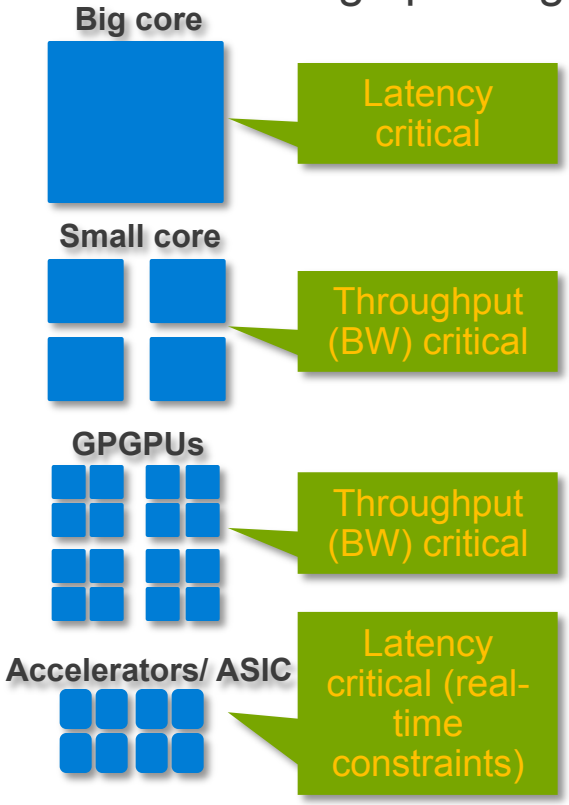
- A NoC paradigm based on top-down approach (application demand/ requirement analysis)

- An efficient design paradigm for future heterogeneous multicores

Summary

• A NoC paradigm based on top-down approach (application demand/ requirement analysis)

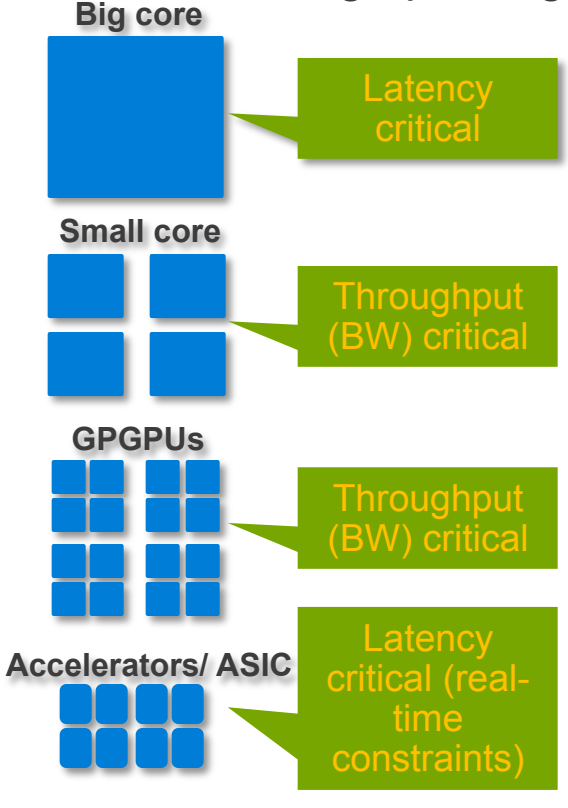
• An efficient design paradigm for future heterogeneous multicores



Summary

• A NoC paradigm based on top-down approach (application demand/ requirement analysis)

• An efficient design paradigm for future heterogeneous multicores



Providing all these guarantees in one network is hard

Multiple networks: each customized for one metric

Summary

• A NoC paradigm based on top-down approach (application demand/ requirement analysis)

• An efficient design paradigm for future heterogeneous multicore m/c