UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

TOWARDS CACHE-COHERENT CHIPLET-BASED ARCHITECTURES WITH WIRELESS INTERCONNECTS

NITISH ARYA

Thesis supervisor: SERGI ABADAL CAVALLÉ (Department of Computer Architecture)

Thesis co-supervisor: ABHIJIT DAS

Degree: Master Degree in Innovation and Research in Informatics (High Performance Computing)

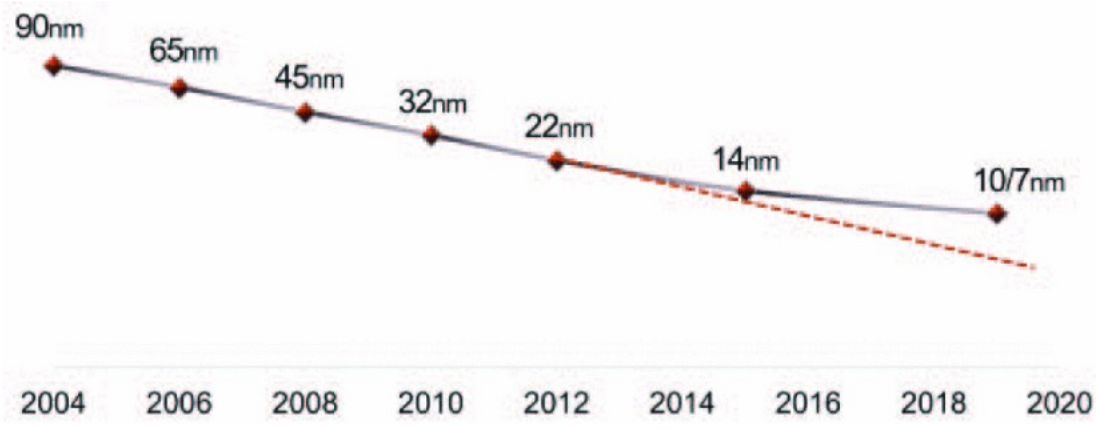
Thesis defence

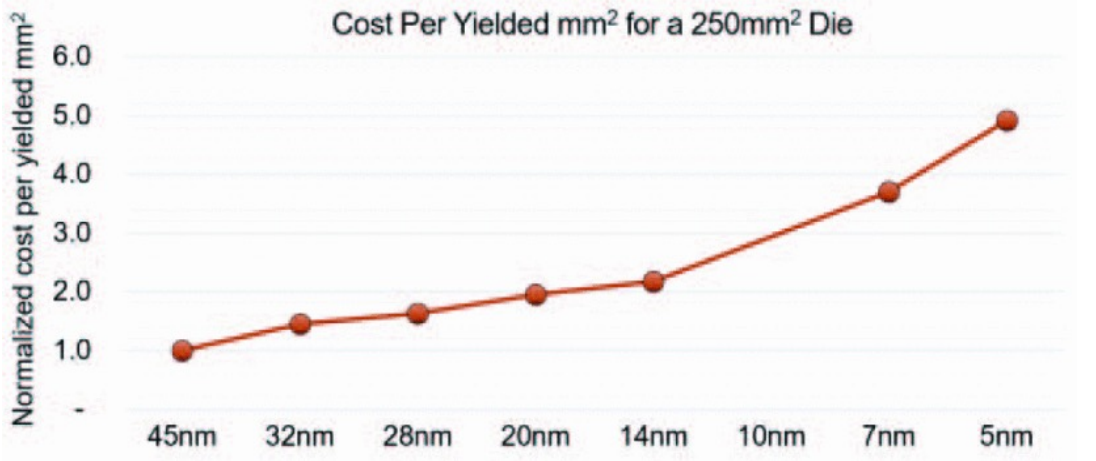
Contents

- Introduction
- Background and motivation
- Methodology
- Results
- Conclusion

Introduction

- High demanding applications; need of more compute power
- Breaking down of Moore's Law
- Die size reaching limits & rising costs





Images source: S. Naffziger et al., "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product," 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 2021, pp. 57-70, doi: 10.1109/ISCA52012.2021.00014.

Chiplets

- Multi core is no longer scalable. Cost is high
- Monolithic design vs Chiplets
- Low manufacturing costs
  - Scaling can continue
  - Customizability

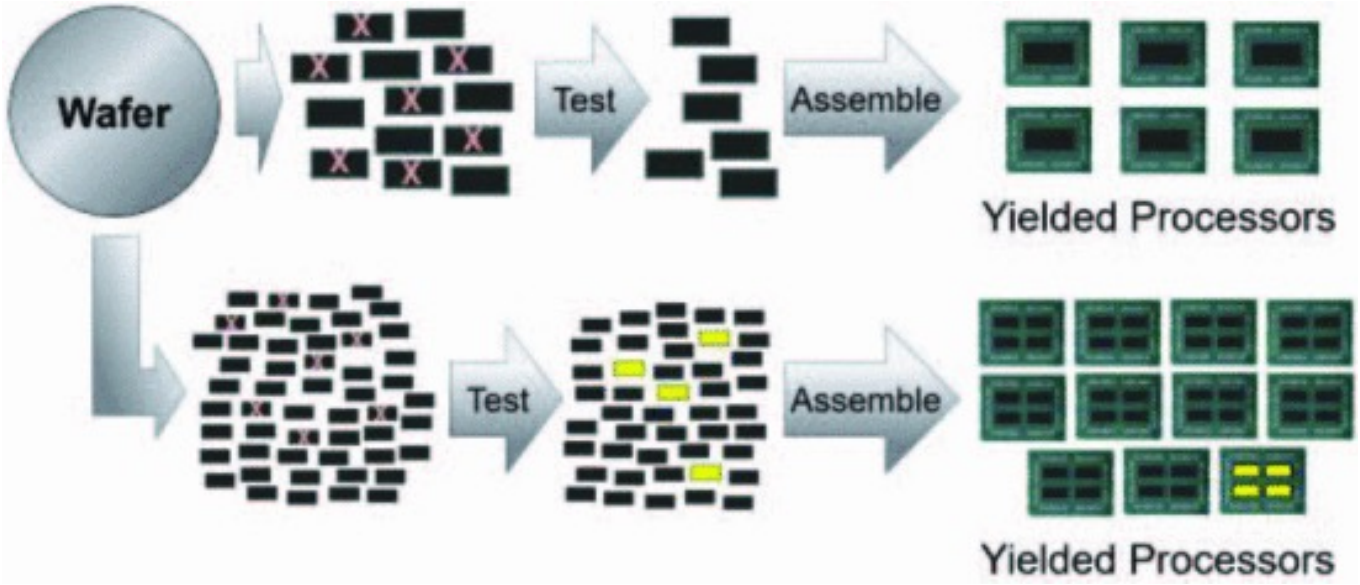


Image source: S. Naffziger et al., "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product," 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 2021, pp. 57-70, doi: 10.1109/ISCA52012.2021.00014.

Introduction

- Heterogenous vs Homogeneous components

- Coherency between components
  - Hardware coherence between the components is necessary
  - Hybrid protocols have been suggested to avoid the cost of debugging complex protocols

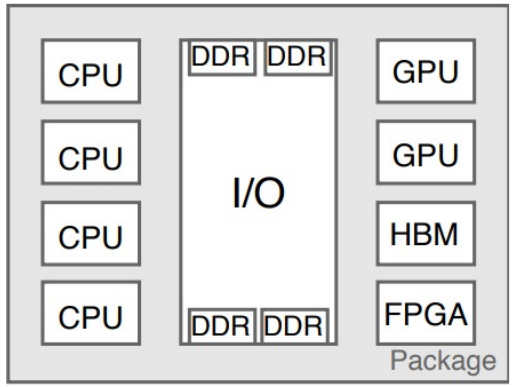


Image source: Pouya Fotouhi, Sebastian Werner, Jason Lowe-Power, and S. J. Ben Yoo. 2019. Enabling scalable chiplet-based uniform memory architectures with silicon photonics. In Proceedings of the International Symposium on Memory Systems (MEMSYS '19). Association for Computing Machinery, New York, NY, USA, 222–334. https://doi-org.recursos.biblioteca.upc.edu/10.1145/3357526.3357564, 2nd image: https://doi.org/10.48550/arXiv.2002.03944

Complexities in chiplet-based approach

- More engineering and more choices
  - Constraint: combinations of partitioning
  - Memory management
    - How coherency should be?
- Chiplet integration and complexity
  - CDCs, SerDes → more area than monolithic
  - **Requires new inter-chiplet communication path**

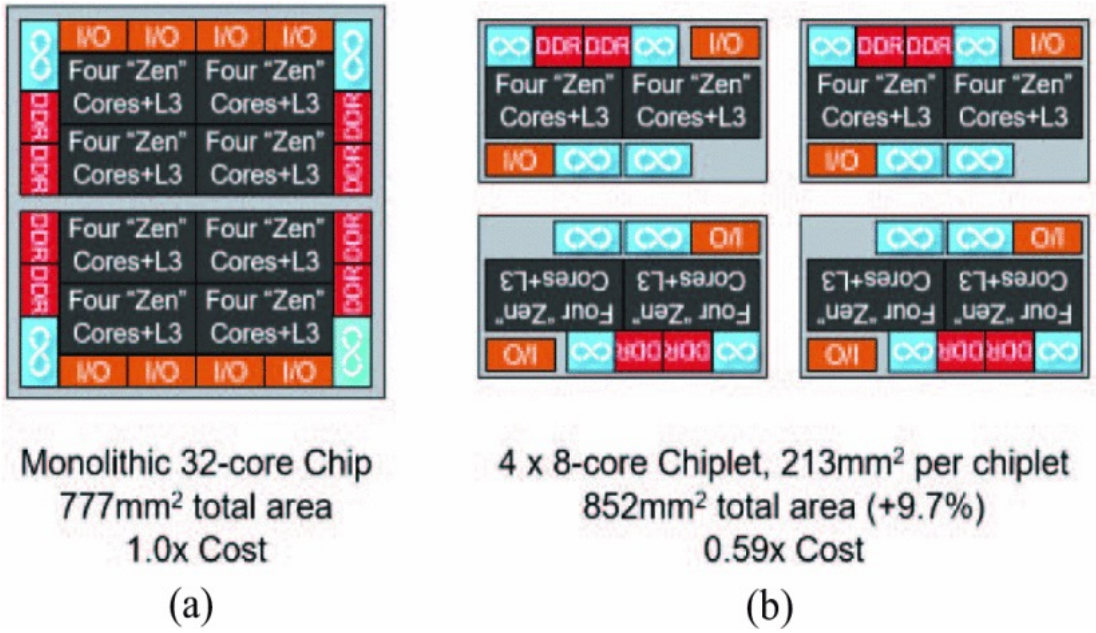


Image source: S. Naffziger et al., "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product," 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 2021, pp. 57-70, doi: 10.1109/ISCA52012.2021.00014.

Many-core Interconnection

- Connecting many core designs
  - Network on Chip(NoC)
- Replaced bus-based and crossbar
- Topology of routers connected point to point

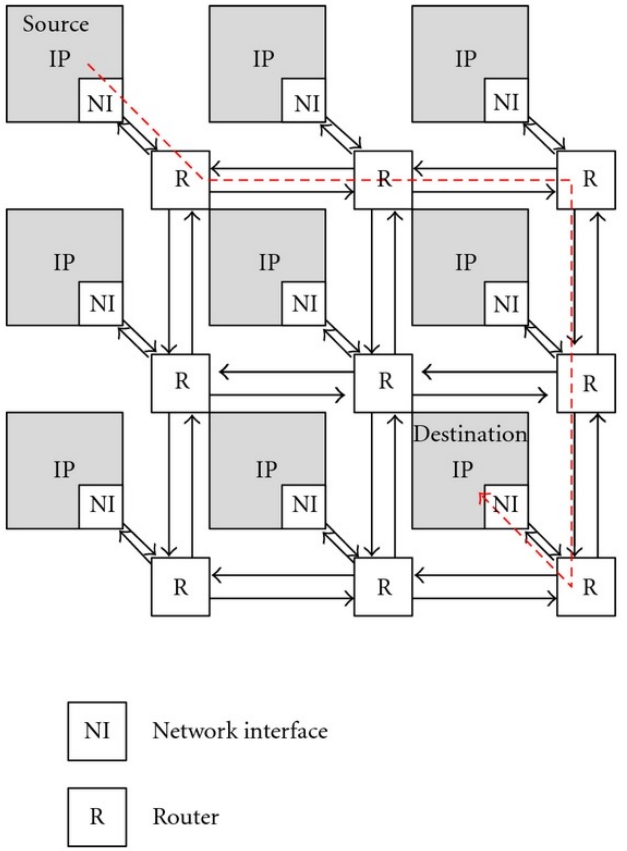


Image source: Wen-Chung Tsai, Ying-Cherng Lan, Yu-Hen Hu, Sao-Jie Chen, "Networks on Chips: Structure and Design Methodologies", *Journal of Electrical and Computer Engineering*, vol. 2012, Article ID 509465, 15 pages, 2012. https://doi.org/10.1155/2012/509465

Inter-chiplet interconnection

- Homogenous and heterogeneous integration – Network on Package
- More and more components
- Physical characteristics cause scaling restriction
- Increase in latency and power consumption

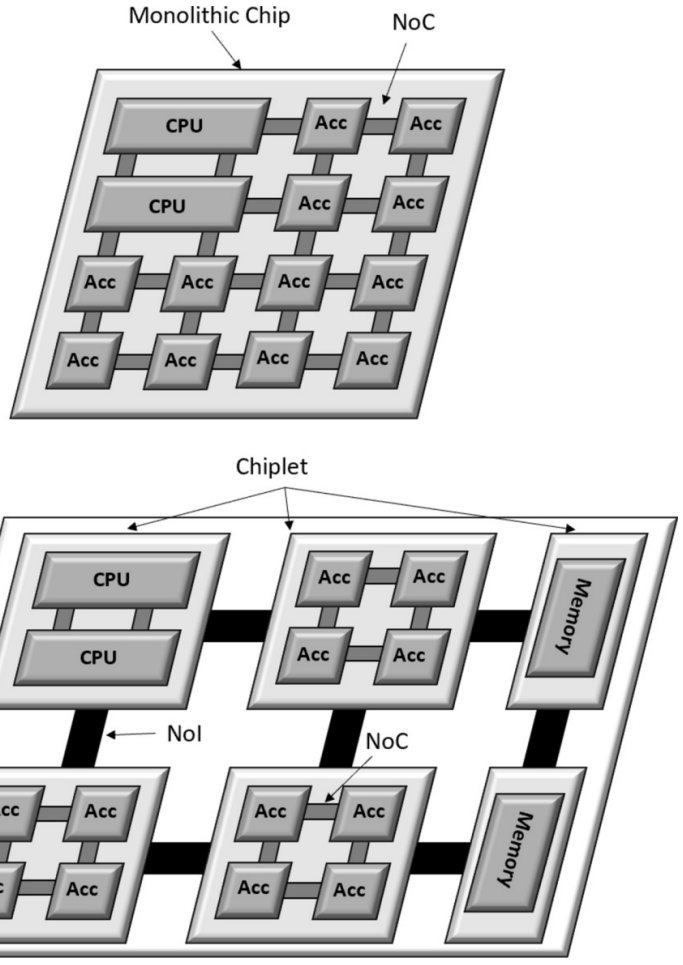


Image source: Ma, X., Wang, Y., Wang, Y. *et al.* Survey on chiplets: interface, interconnect and integration methodology. *CCF Trans. HPC* **4**, 43–52 (2022). https://doi.org/10.1007/s42514-022-00093-0

Alternatives to just wires

- Inter-chip photonics
  - Optical networks
  - Designing is complex
- 3D ICs
  - Stacking and Intergration complexity
  - Layer stacking → heat density
- RF Interconnects
  - Similar problems to photonics
  - Still guiding is wired!
  - Interferences

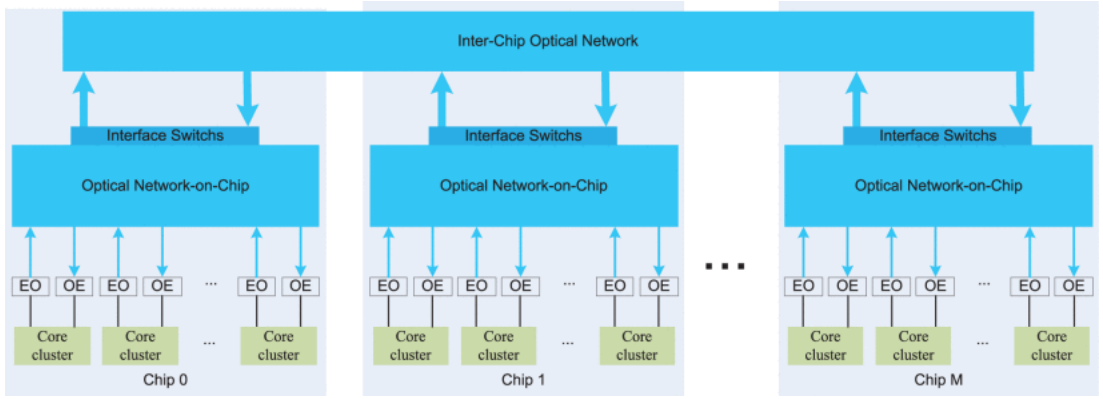


Image source: X. Wu et al., "UNION: A Unified Inter/Intrachip Optical Network for Chip Multiprocessors," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 22, no. 5, pp. 1082-1095, May 2014, doi: 10.1109/TVLSI.2013.2263397.

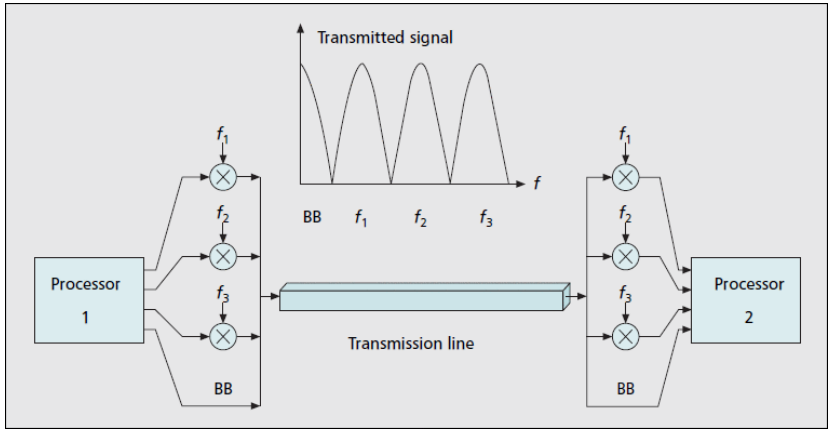


Image source: E. Socher and M. -C. F. Chang, "Can RF Help CMOS Processors? [Topics in Circuits for Communications]," in IEEE Communications Magazine, vol. 45, no. 8, pp. 104-111, August 2007, doi: 10.1109/MCOM.2007.4290322.

Wireless Interconnects

- Transmission distance independent
- Reconfigurable during design
  - Logical topology modification without touching physical topology
  - Coherency modification is not required
- Broadcast, multicast efficient
- Improved scalability
  - Low latency and energy consumption
  - Increased thorughput

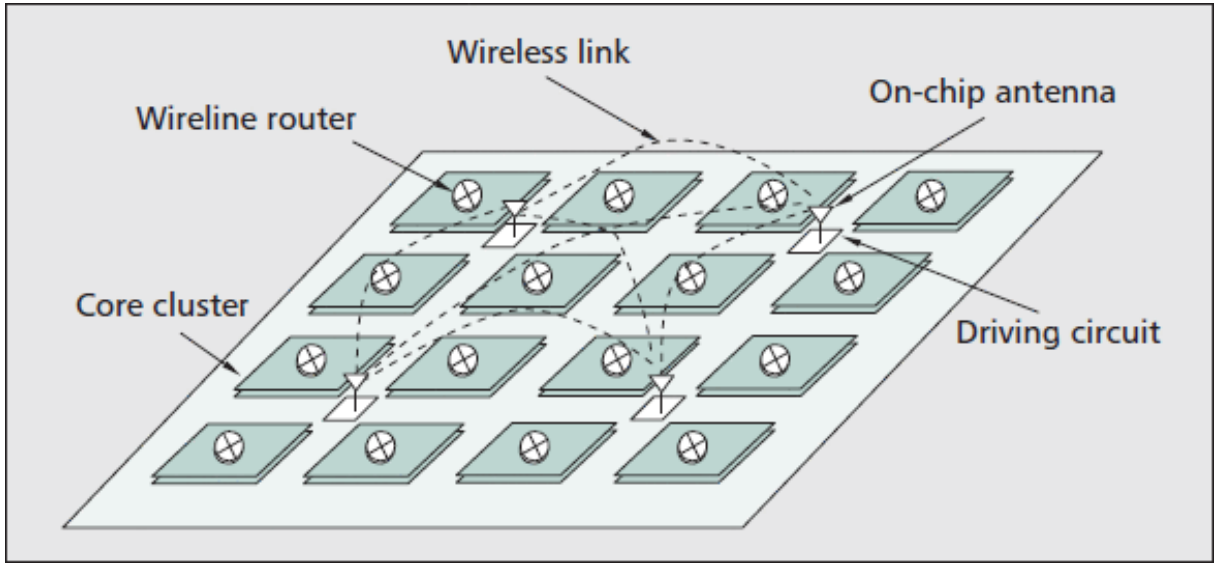
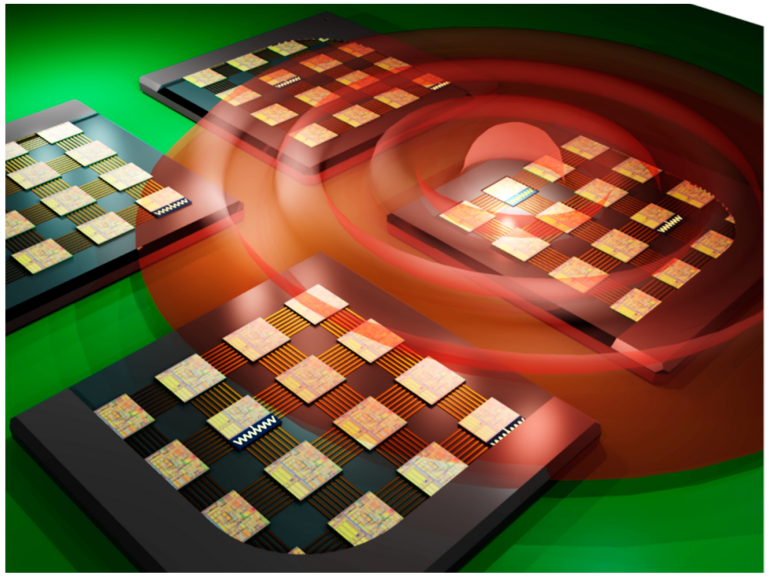


Image source: S. Abadal, E. Alarcón, A. Cabellos-Aparicio, M. C. Lemme and M. Nemirovsky, "Graphene-enabled wireless communication for massive multicore architectures," in IEEE Communications Magazine, vol. 51, no. 11, pp. 137-143, November 2013, doi: 10.1109/MCOM.2013.6658665.

Literature Review – Related work

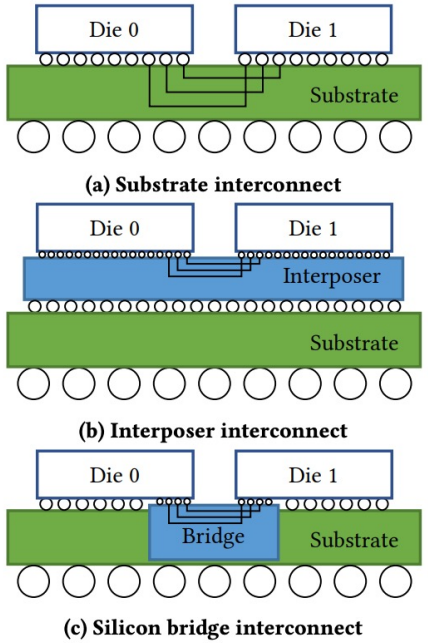
- "The Advances, Challenges and Future Possibilities of Millimeter-Wave Chip-to-Chip Interconnections for Multi-Chip Systems." *
  - Covers every aspect from network to physical layer
  - Inter-chip system has both CPU and GPU
- Traffic characterization is not a part



* Ganguly, A.; Ahmed, M.M.; Singh Narde, R.; Vashist, A.; Shamim, M.S.; Mansoor, N.; Shinde, T.; Subramaniam, S.; Saxena, S.; Venkataraman, J.; et al. The Advances, Challenges and Future Possibilities of Millimeter-Wave Chip-to-Chip Interconnections for Multi-Chip Systems. *J. Low Power Electron. Appl.* **2018**, *8*, 5. https://doi.org/10.3390/jlpea8010005

Literature Review – Related work

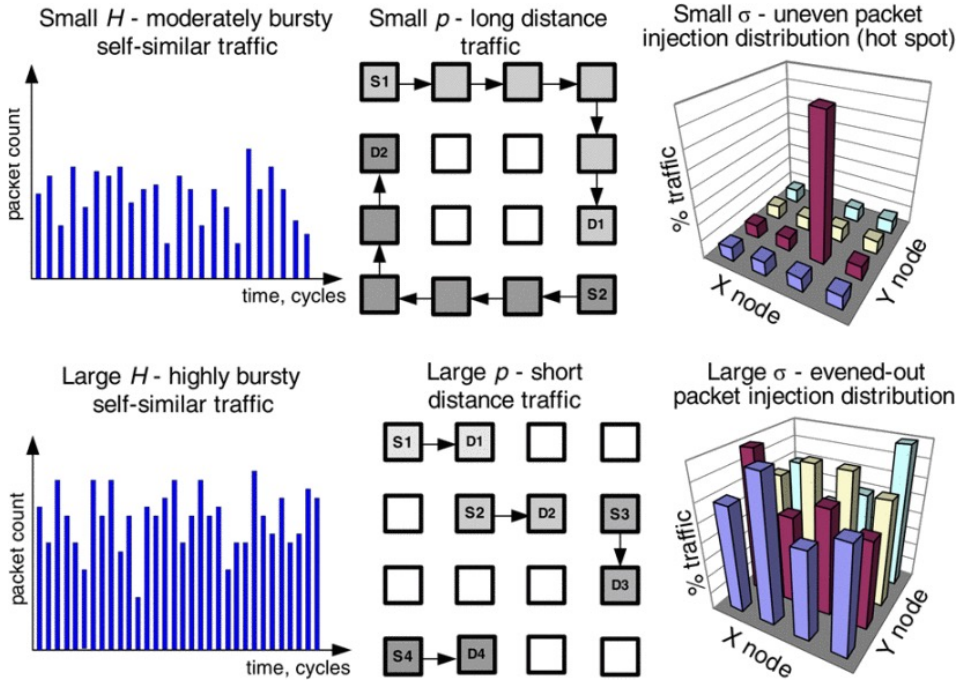
- "Seizing the Bandwidth Scaling of On-Package Interconnect in a Post-Moore's Law World" *
- Provides a cache coherence protocol to exploit interconnect utilization for better performance
- Still wired and designs a new protocol



* Grigory Chirkov and David Wentzlaff. 2023. Seizing the Bandwidth Scaling of On-Package Interconnect in a Post-Moore's Law World. In Proceedings of the 37th International Conference on Supercomputing (ICS '23). Association for Computing Machinery, New York, NY, USA, 410–422. https://doi-org.recursos.biblioteca.upc.edu/10.1145/3577193.3593702

Literature Review – Related work

- "A Statistical Traffic Model for On-Chip Interconnection Networks" *
- Presents a comprehensive traffic model for NoCs
- Basis of our work
  - but for chiplets
  - Keeping in mind wireless integration



* V. Soteriou, Hangsheng Wang and L. Peh, "A Statistical Traffic Model for On-Chip Interconnection Networks," 14th IEEE International Symposium on Modeling, Analysis, and Simulation, Monterey, CA, USA, 2006, pp. 104-116, doi: 10.1109/MASCOTS.2006.9.

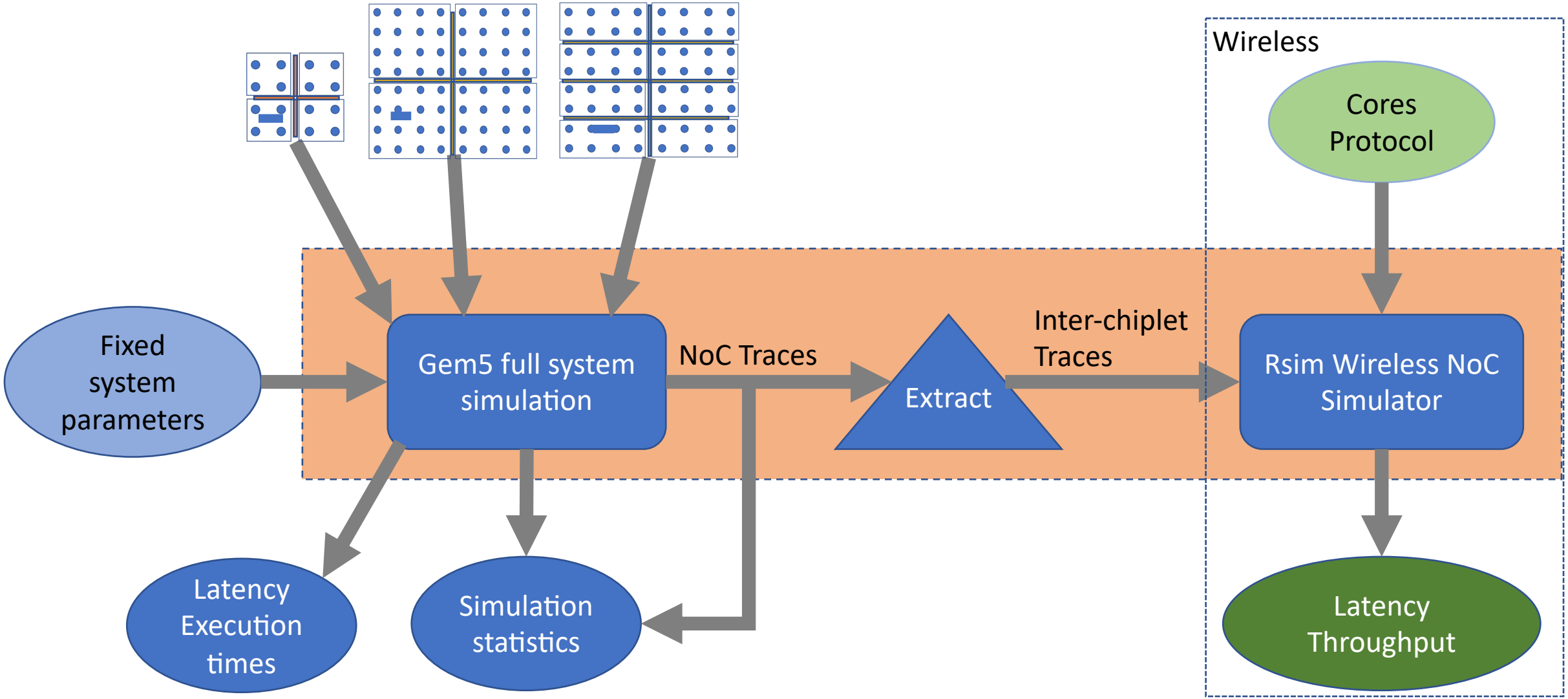
Motivation

- Addressing the challenges associated with interconnecting chiplets
- More components mean more communication
  - Maintain scalability while keeping area, power and latency optimal
- Current hybrid interconnection technologies have their challenges
- Wireless interconnects is a good candidate as it does not try to modify the cache coherence protocol or design a new one

Objectives

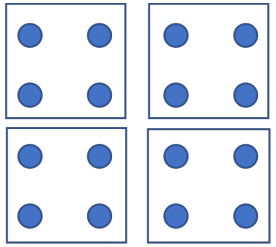
- Characterize wired interconnect communication traffic for chiplet-based architectures
  - Study packet distribution, latency and burstiness
- Investigate potential areas where wireless could benefit specifically in terms of latency and without modifying the coherence protocol
  - Find a hybrid interconnect design
- Assess advantages and trade-offs and understand limitation and challenges

Simulation flow

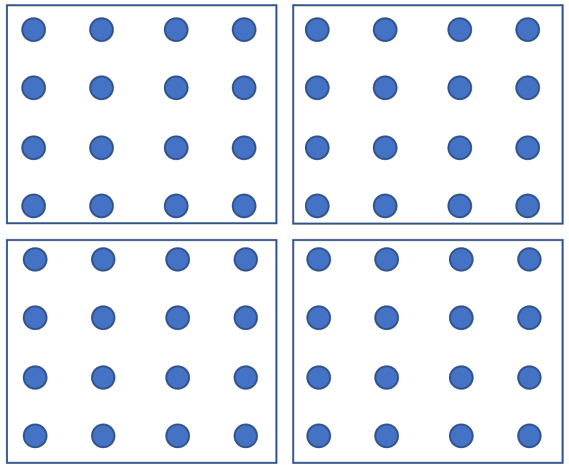


Methodology

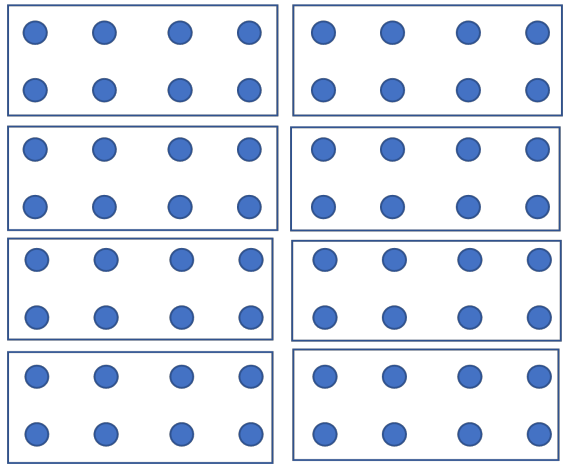
- Designing the experiment
  - Three chiplet-based architectures
  - Baseline is the monolithic 16 core system



16 cores with 4 chiplets

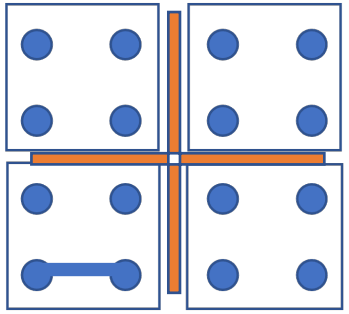


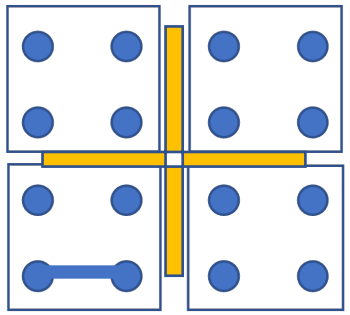
64 cores with 4 chiplets



64 cores with 8 chiplets

Latencies

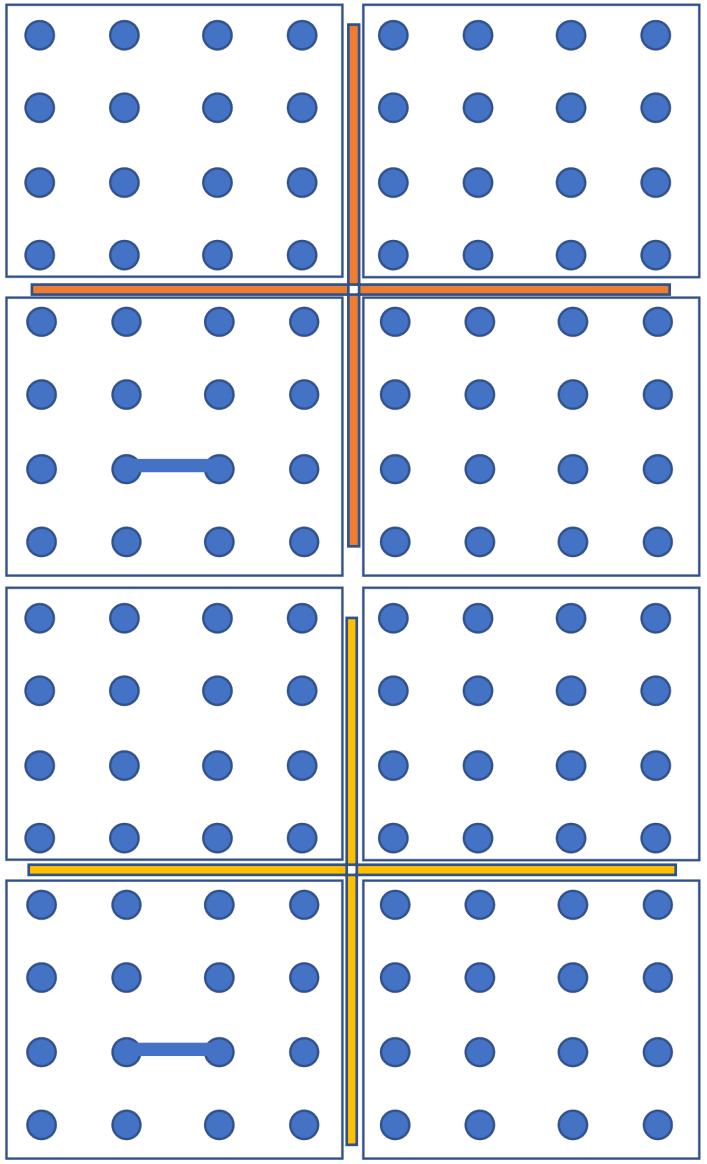


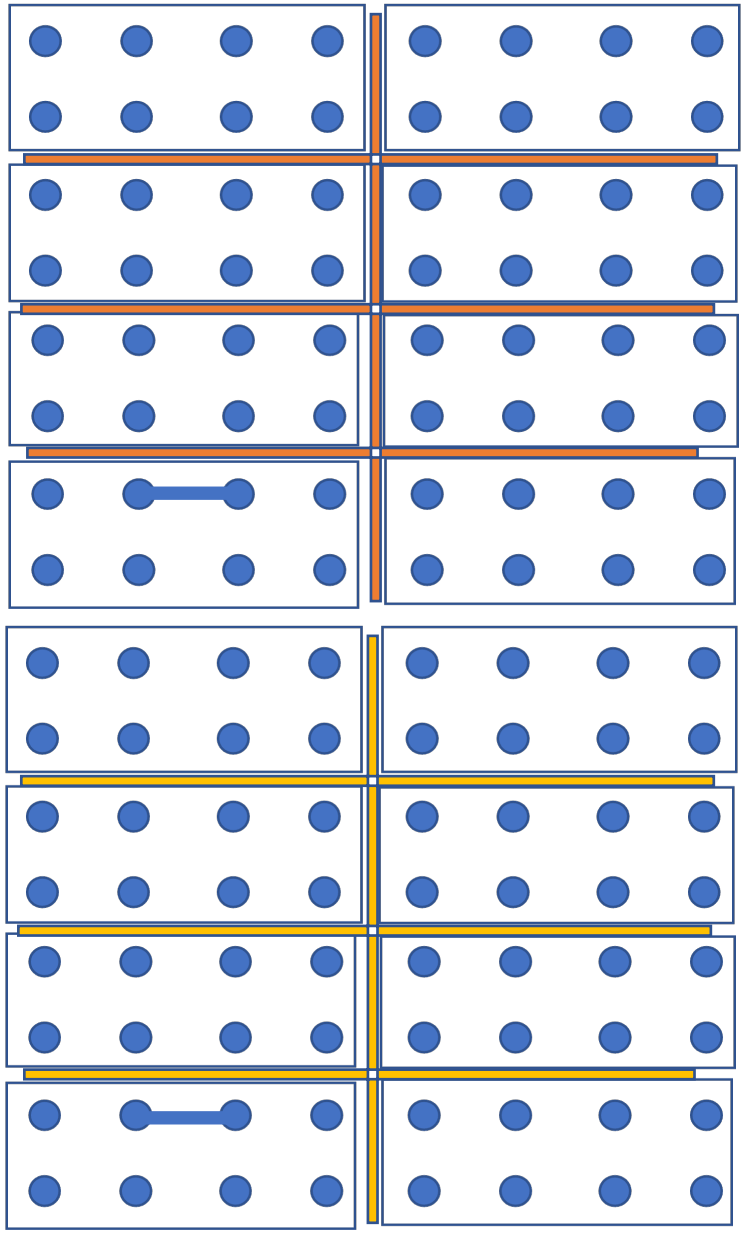


Intra-chiplet link

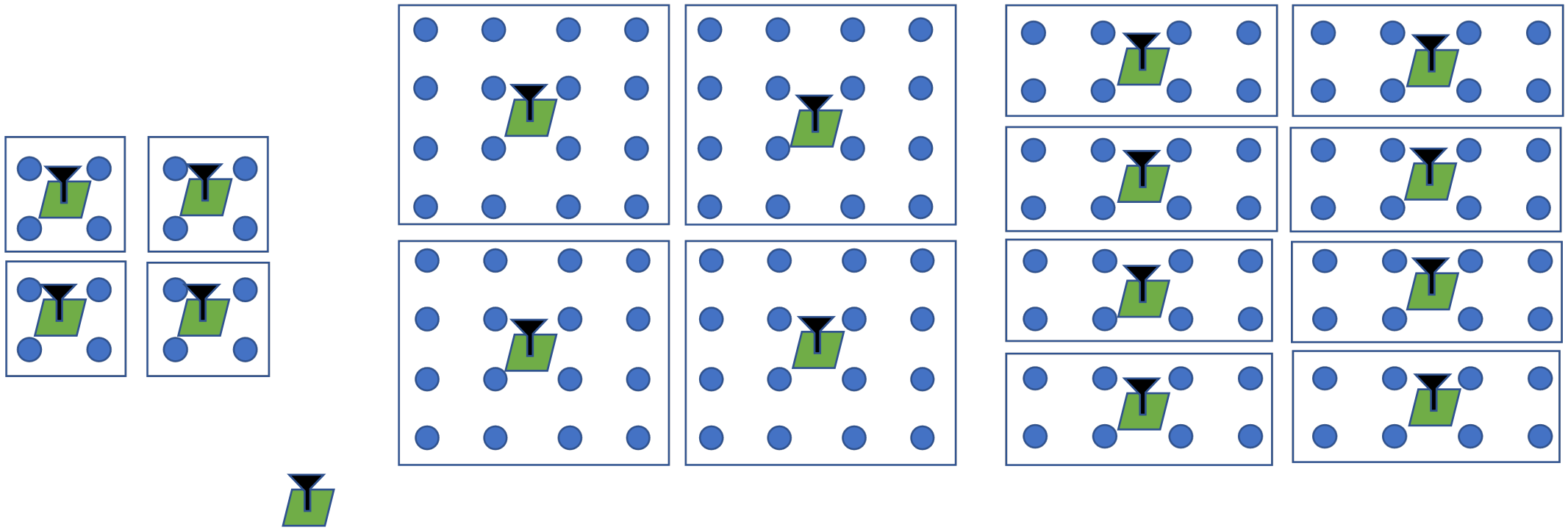
10x

100x





Wireless links



On-package antenna

PARSEC

- All the 8 applications were run with **simsmall** input set
- *parsecmgmt* was the common utility for managing
- Some applications either produced an error(x264) or did not finish in time(streamcluster)

Table 3.2: Simulated Benchmarks

Benchmark	Simulation time
blackscholes	small
bodytrack	medium
canneal	large
dedup	large
ferret	medium
fluidanimate	large
freqmine	medium
vips	large

https://parsec.cs.princeton.edu/index.htm

Simulators

- We use gem5, which is quasi-cycle accurate
- Modular – components for memory, CPU…
- Allows full-system simulation with modifying the linux kernel
- The interconnection network is Garnet

Nathan Binkert et al. "The Gem5 Simulator". In: SIGARCH Comput. Archit. News 39.2 (Aug. 2011), pp. 1–7. issn: 0163-5964. doi: 10.1145/ 2024716.2024718. url: https://doi-org.recursos.biblioteca.upc.edu/10.1145/2024716.2024718.

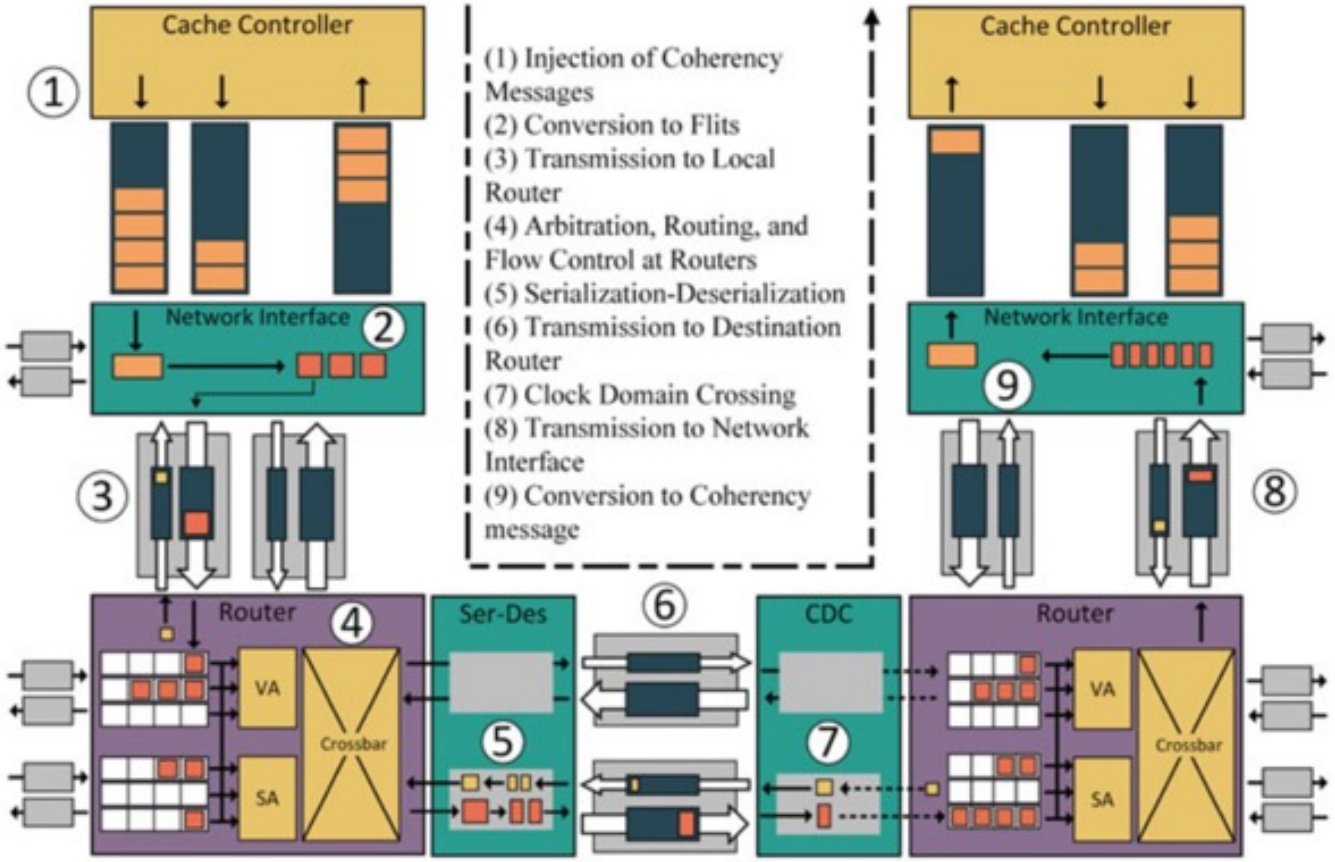
Srikant Bharadwaj et al. "Kite: A Family of Heterogeneous Interposer Topologies Enabled via Accurate Interconnect Modeling". In: 2020 57th ACM/IEEE Design Automation Conference (DAC). 2020, pp. 1– 6. doi: 10.1109/DAC18072.2020.9218539.

Fixed parameters for gem5

Table 3.3: Fixed System Parameters

Parameter	Value
ISA	x86
CPU Type	TimingSimpleCPU
Simulation mode	Full System
L1[data/inst.] size, associativity	32kB, 4
L2 size, associativity	256kB, 8
Cache coherency	MESI_Two_Level
Kernel	x86-linux-kernel-4.19.83
OS	Ubuntu 18.04.2 LTS pre-loaded with PARSEC benchmark
Memory	512MB
Clock	1GHz
Interconnection Network	Garnet 3.0(HeteroGarnet)

Obtaining gem5 communication traces



Lifecycle of a coherence message in heterogarnet

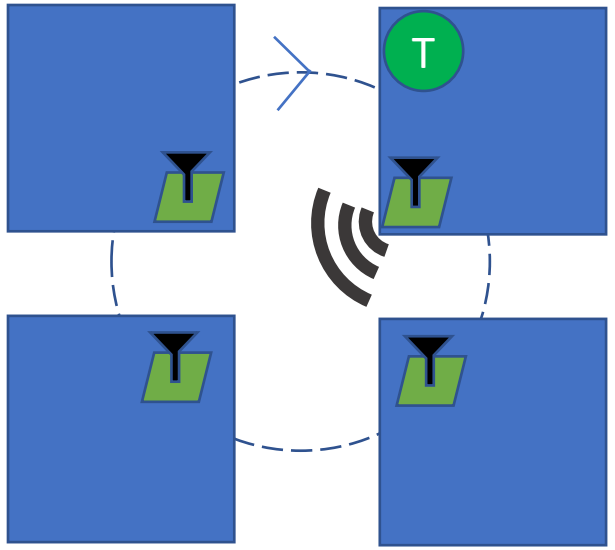
https://www.gem5.org/documentation/general_docs/ruby/heterogarnet/

RSim

- In-house wireless network simulator for studying simulating wireless interconnect traffic
- Input trace: **cycle, source node**
  - Modified to also take the size of message as third field
- Different MAC protocols supported
  - TOKEN, FuzzyToken, BRS…
- Outputs
  - Latencies per node and average arithmetic/geometric mean
  - Throughput

TOKEN Protocol

- Used in telecommunication
  - Already implemented
- Only device(chiplet) with token can transmit
  - No collisions
  - Ring network – fair access
- Modifications to RSim:
  - *Transmission latency is size dependent*

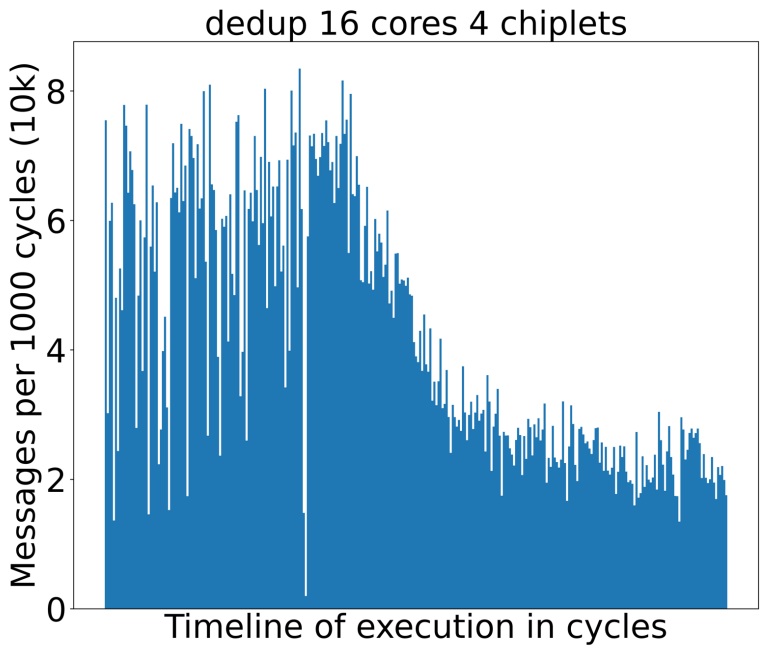


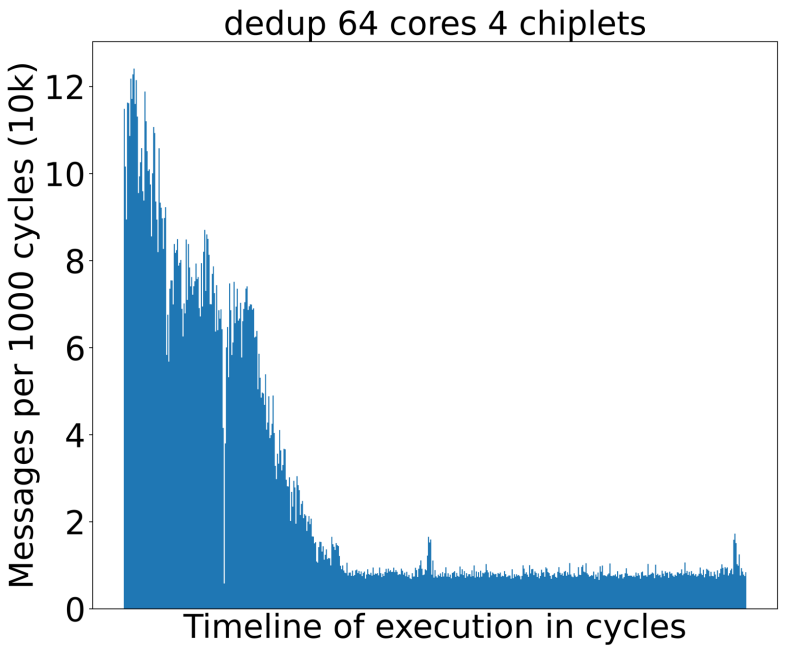
Results

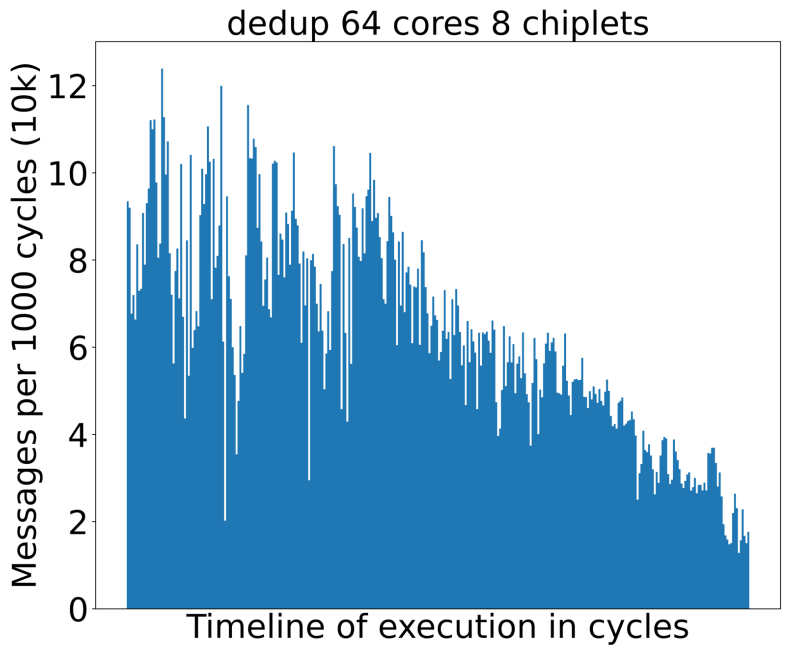
- Characterization
  - Temporal profile
  - Spatial profile
- Performance
  - Execution times
  - Latency
- Wireless

Characterization – Temporal profile

- Inter-chiplet traffic for application dedup showing bursty traffic
- Increase in inter-chiplet messages from 16 to 64 cores

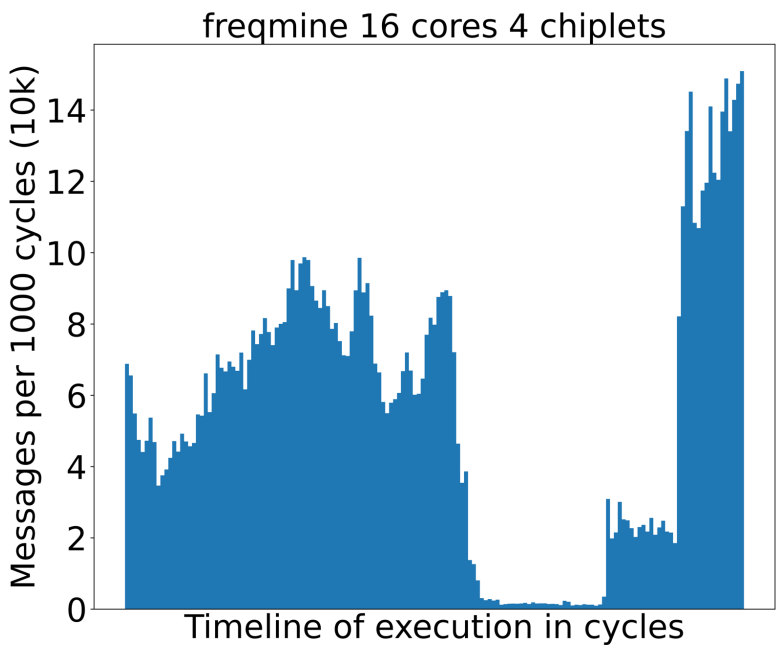


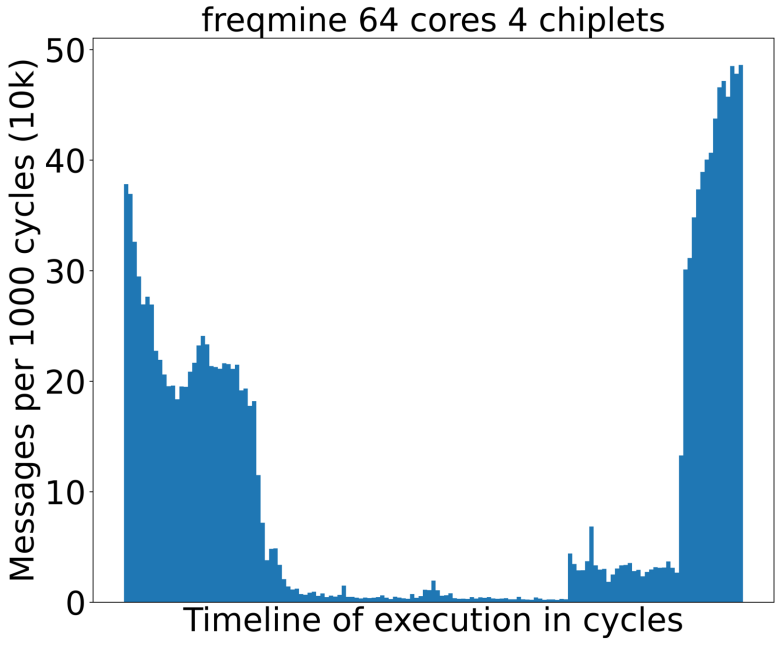


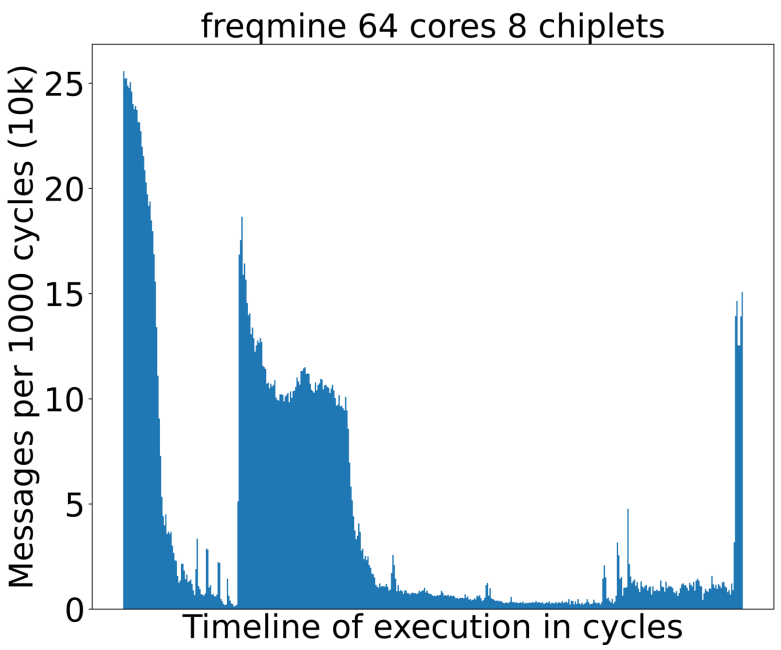


Characterization – Temporal profile

- Inter-chiplet traffic for application freqmine
- The number of messages decrease when chiplets increase

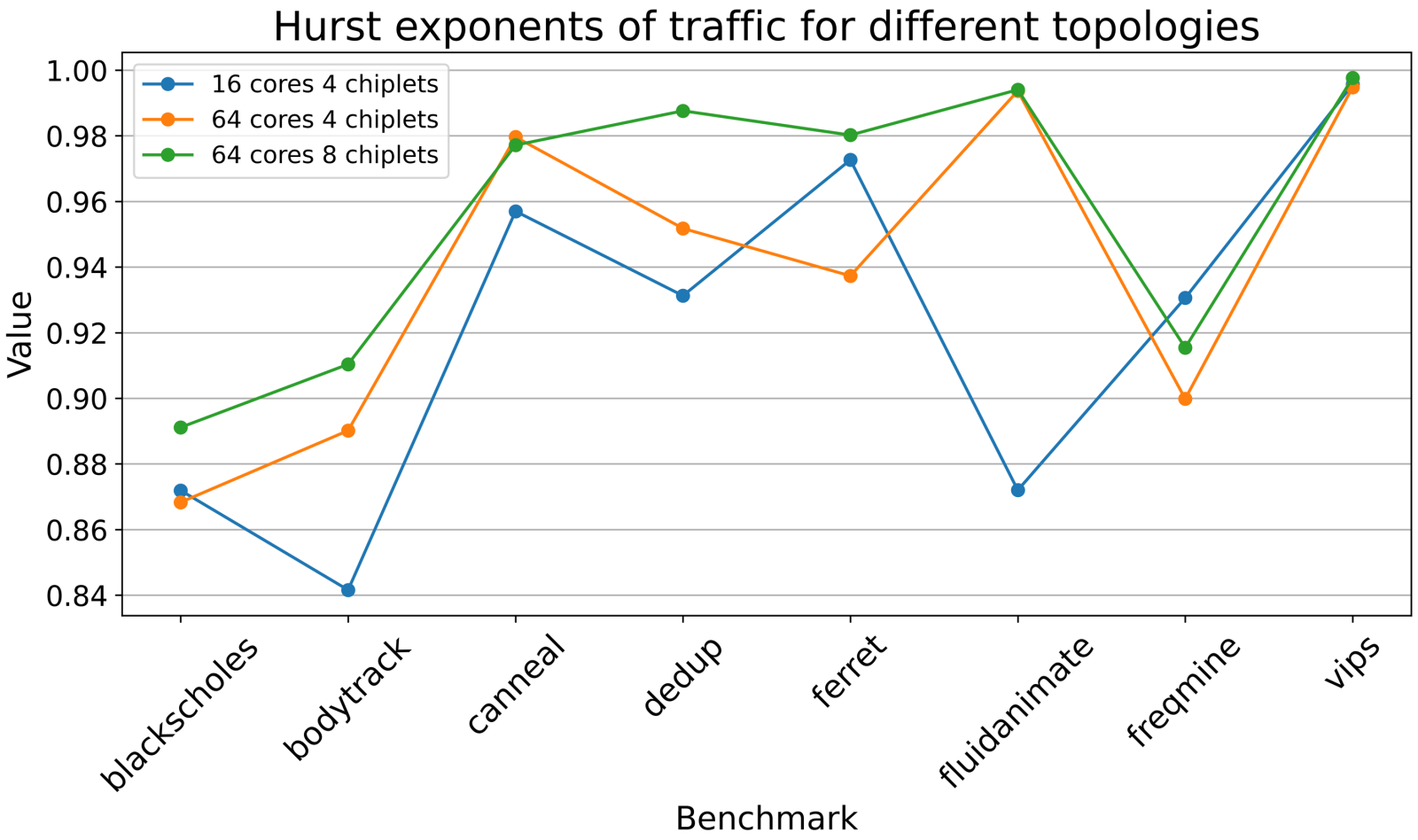






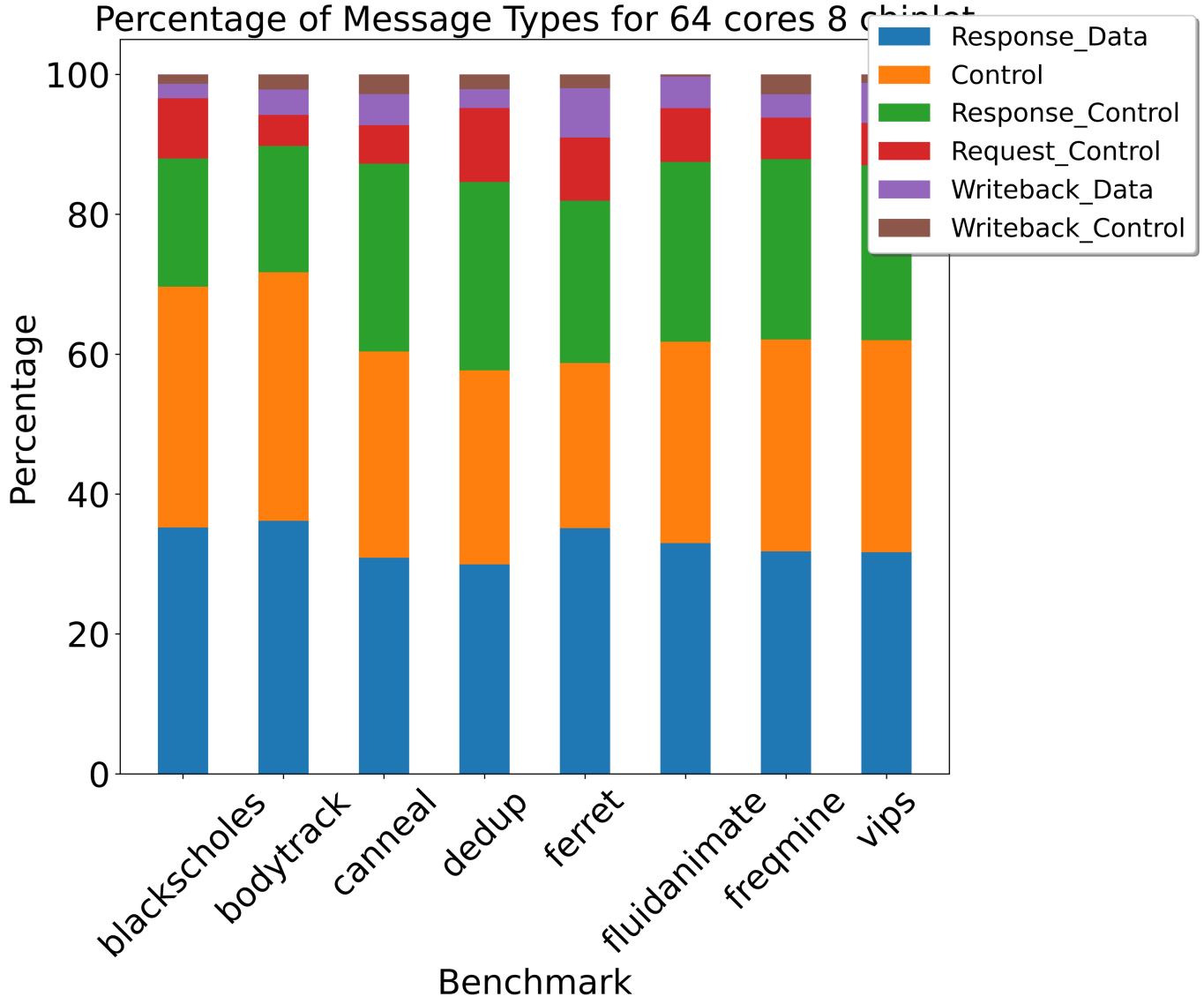
Characterization - Temporal profile

- $0.5 \leq H \leq 1$
- Higher *H* means self-similar bursty traffic
- Generally, 64 core 8 chiplet traffic is the most bursty
- Key points
  - fluidanimate's range
  - freqmine's reverse behaviour



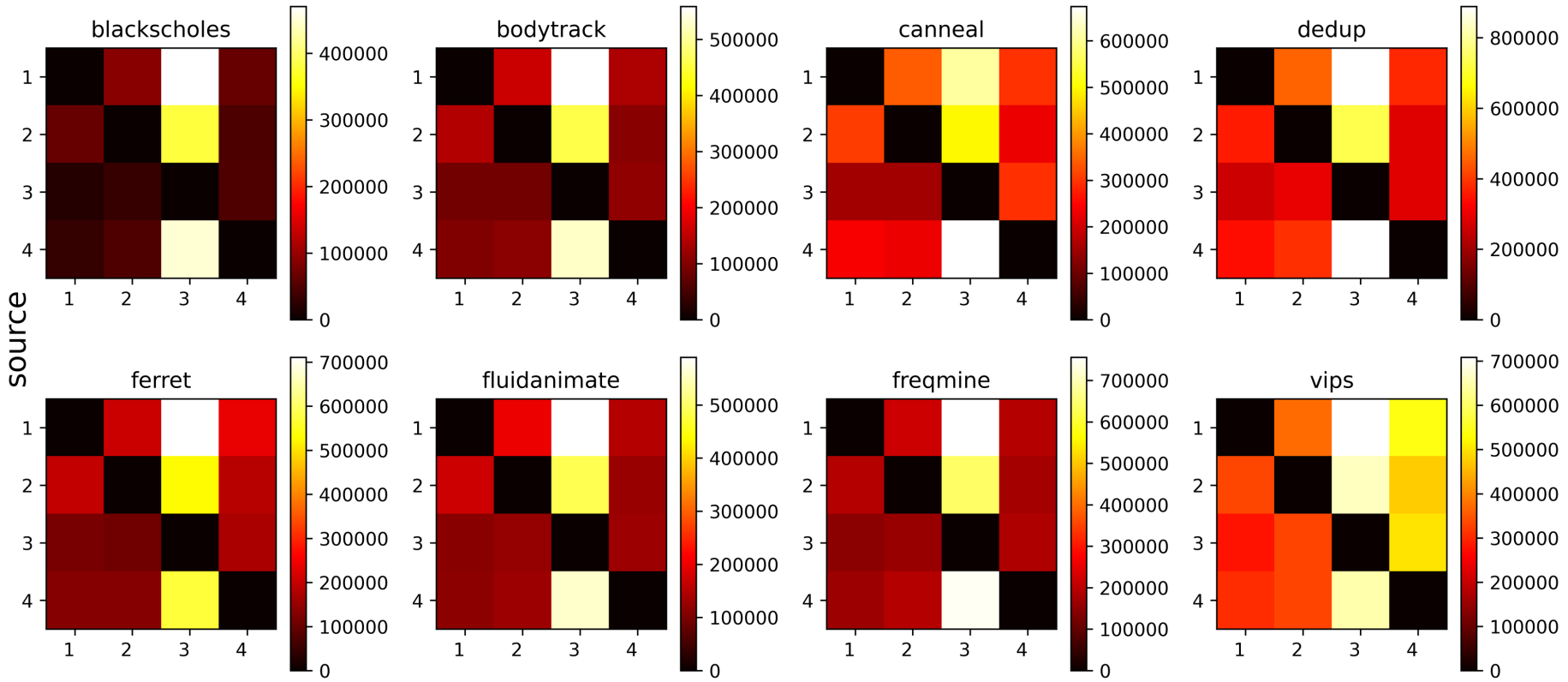
Characterization

- Distribution of type of messages crossing the chiplet boundary
- 60-70% traffic is Control and Response_Data
- The distribution is similar in all the configurations



Characterization – Spatial profile

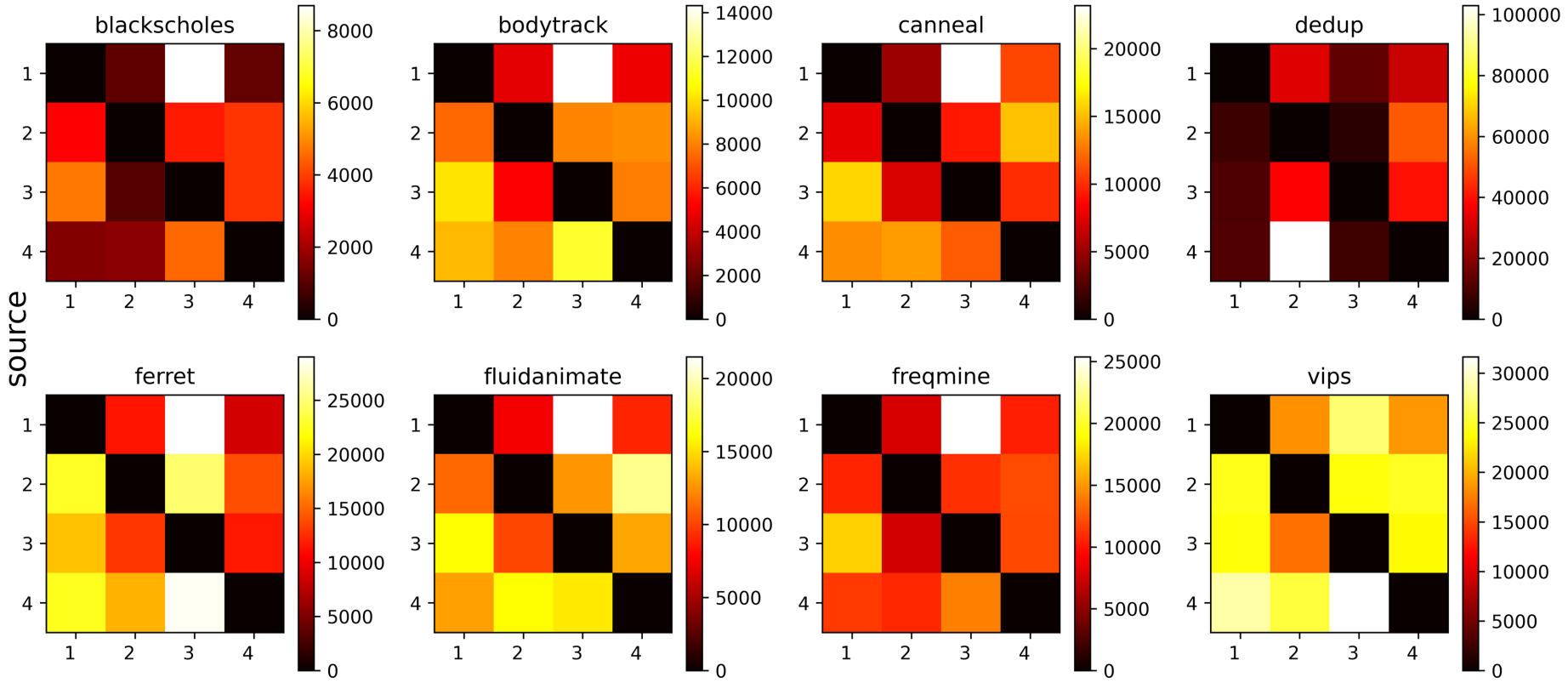
- X-axis are destinations
- Hotspot chiplet is 1 in most cases
- vips, canenal and dedup have relatively more communicative traffic



Characterization – Spatial profile

- Chiplet 1 and 4 can be considered hotspots
- Application behaviour changes from 16 to 64 core scaling
- Less inter-chiplet messages compared to 16 core 4 chiplet

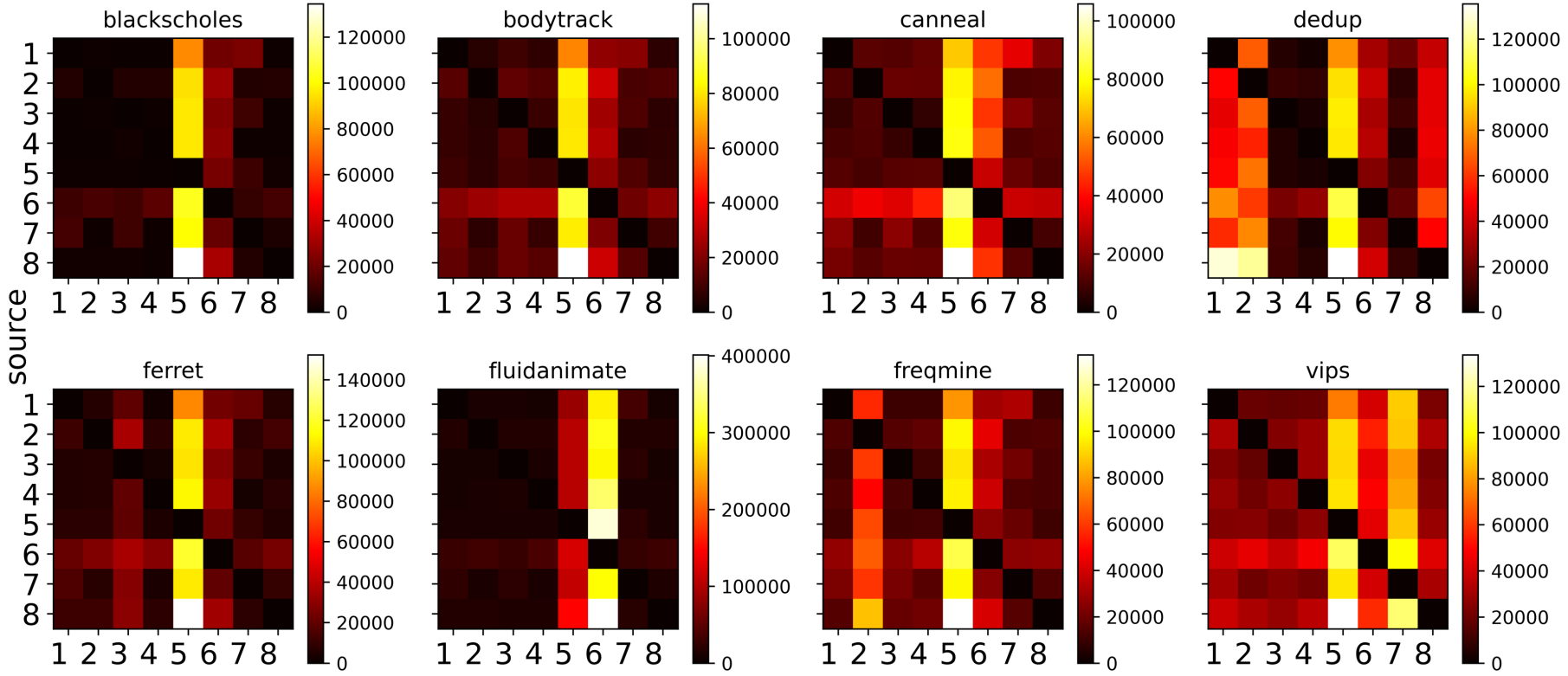
Heatmaps for 64 cores 4 chiplets



Characterization – Spatial profile

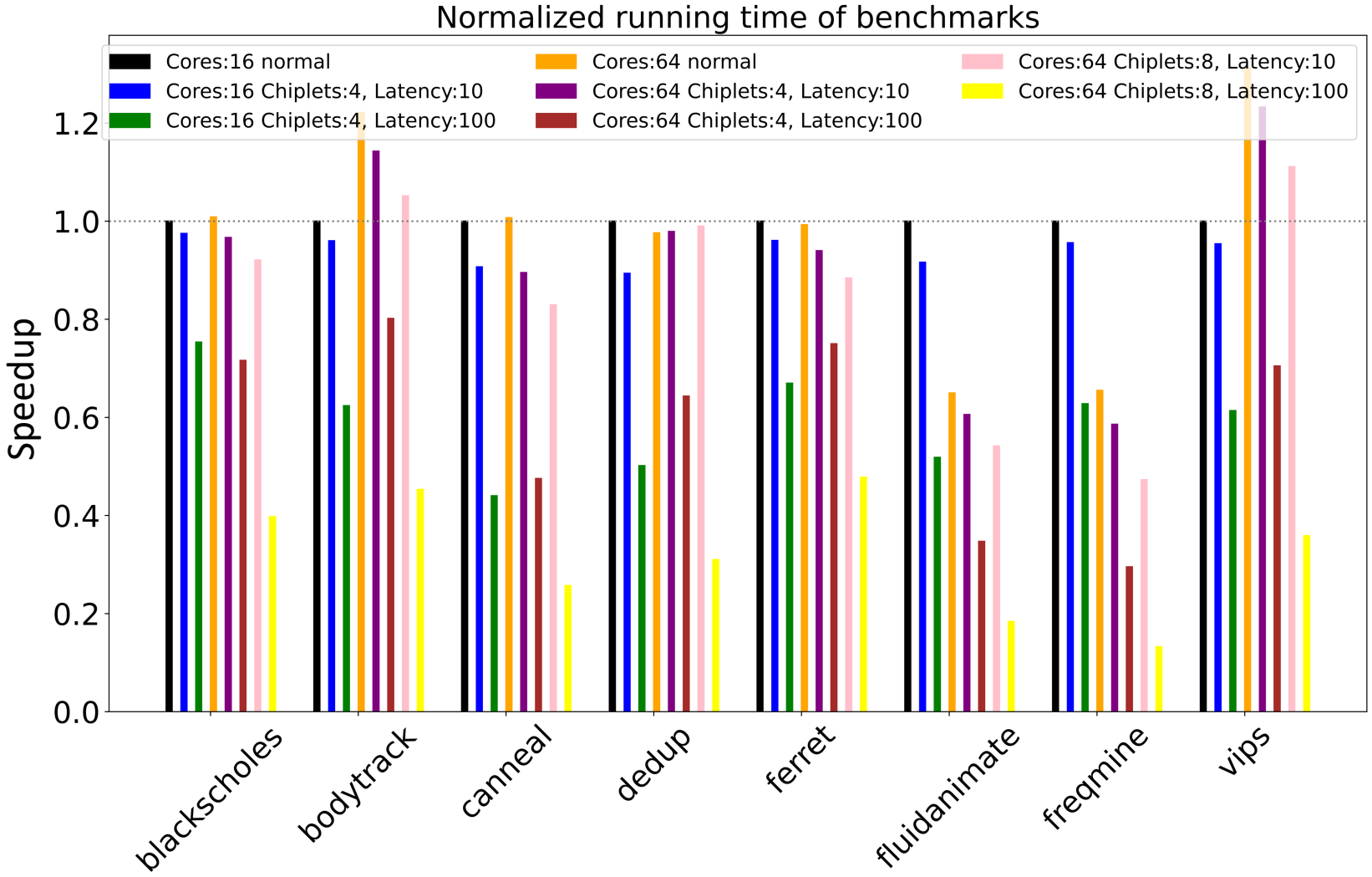
- Chiplet 6 and 8 are hotspots but not very distinctive
- Number of messages increase as chiplets scale
- dedup and vips show communicative intensive pattern

Heatmaps for 64 cores 8 chiplets



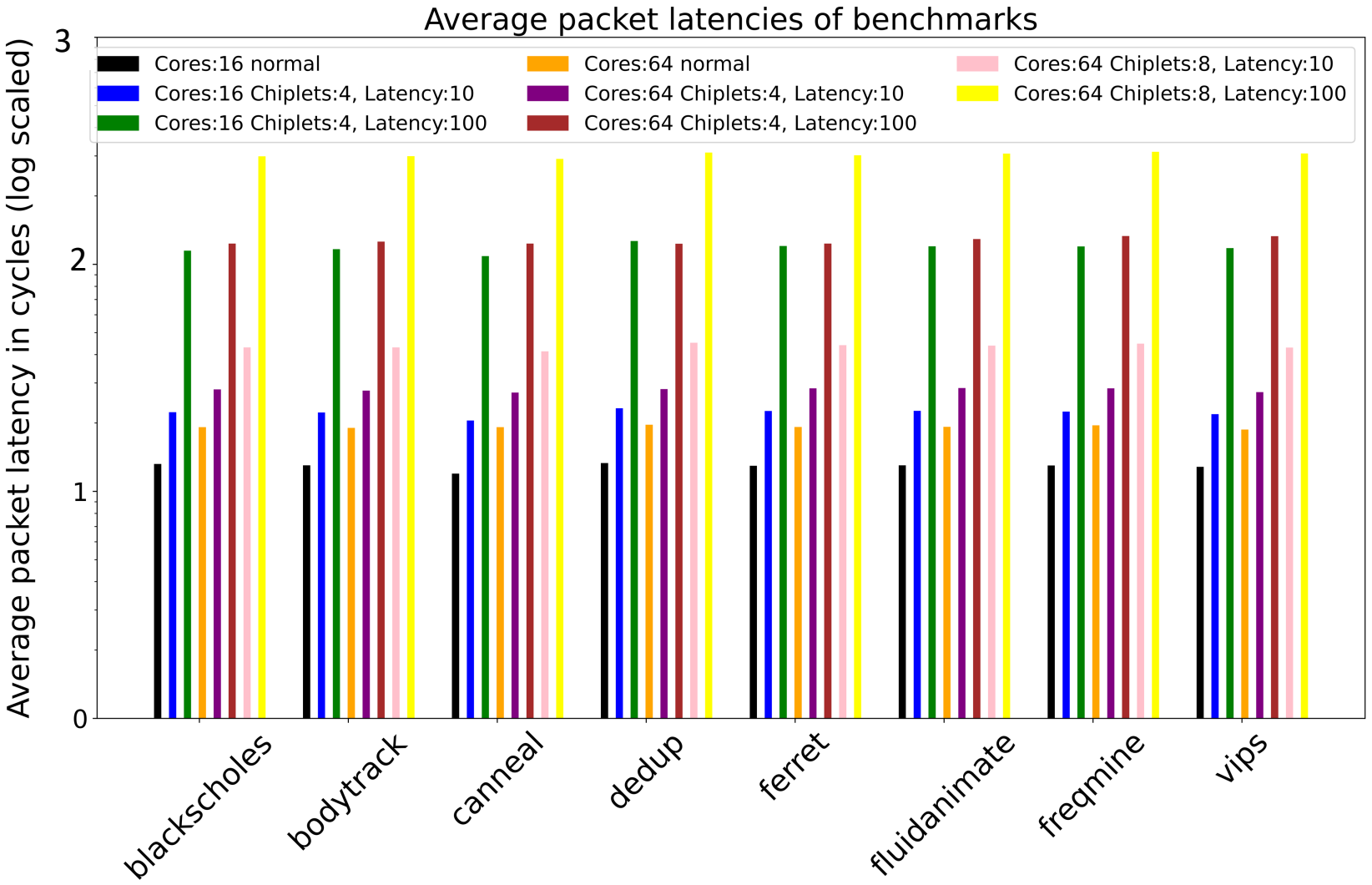
Performance – Execution times

- (brown and yellow) Chiplets indicate poor scaling in terms of performance
  - More chiplets on same base system mean more slowdown
- (green and brown) blackscholes, fluidanimate and freqmine lose performance from 16 to 64 cores under 4 chiplets



Performance- Latencies

- (brown, purple and yellow, pink) Latencies increase when chiplets increase
  - Segmentation of chiplets on the same system also adds to the average latencies, almost twice if chiplets are doubled
- The jump in latency is not quantitatively progressed
  - Increasing inter-chiplet latency by 10/100 times does not mean average packet latency increases 10/100 times
  - This is a potential chance for wireless to cover up the increase in latency



Wireless results

- The latencies listed are excessively high, the reason being the injection rate of the traces and other factors
  - Rsim must be supplied with correct injection rate, bandwidth and buffer size for each of the transmitting nodes
- Experimentation with other MAC protocols will further decrease the average packet latency as token induces waiting periods in certain situations.

Table 4.1: Average packet latencies for inter-chiplet data(cores)

Benchmark	Cycles per Packet(16)	(64)
blackscholes	32.8307	224.296
bodytrack	587.901	170409
canneal	8.74244e+06	1.77433e+07
dedup	1016.42	44710.3
ferret	749.175	28900.1
fluidanimate	324020	229.056
freqmine	65.9824	8.05431e+06
vips	234.271	2.95392e+07

Conclusions

- Around 70% of the traffic for an application is inter-chiplet and the rest is within the chiplet
- Inter-chiplet traffic is highly self similar as the Hurst exponent is higher than 0.8 for all applications
- The spatial profile of the traffic remains dynamic with change in chiplet configuration. The emergence of hotspots is application specific.
- Integrating wireless with the current wired interconnects will not require a change in the cache coherence protocol