Modeling many-core processor interconnect scalability for the evolving performance, power and area relation

David Smelt

June 25, 2018

Supervisor: drs. T.R. Walstra

BSc Computer Science thesis, University of Amsterdam

Table of contents

The need for energy efficiency

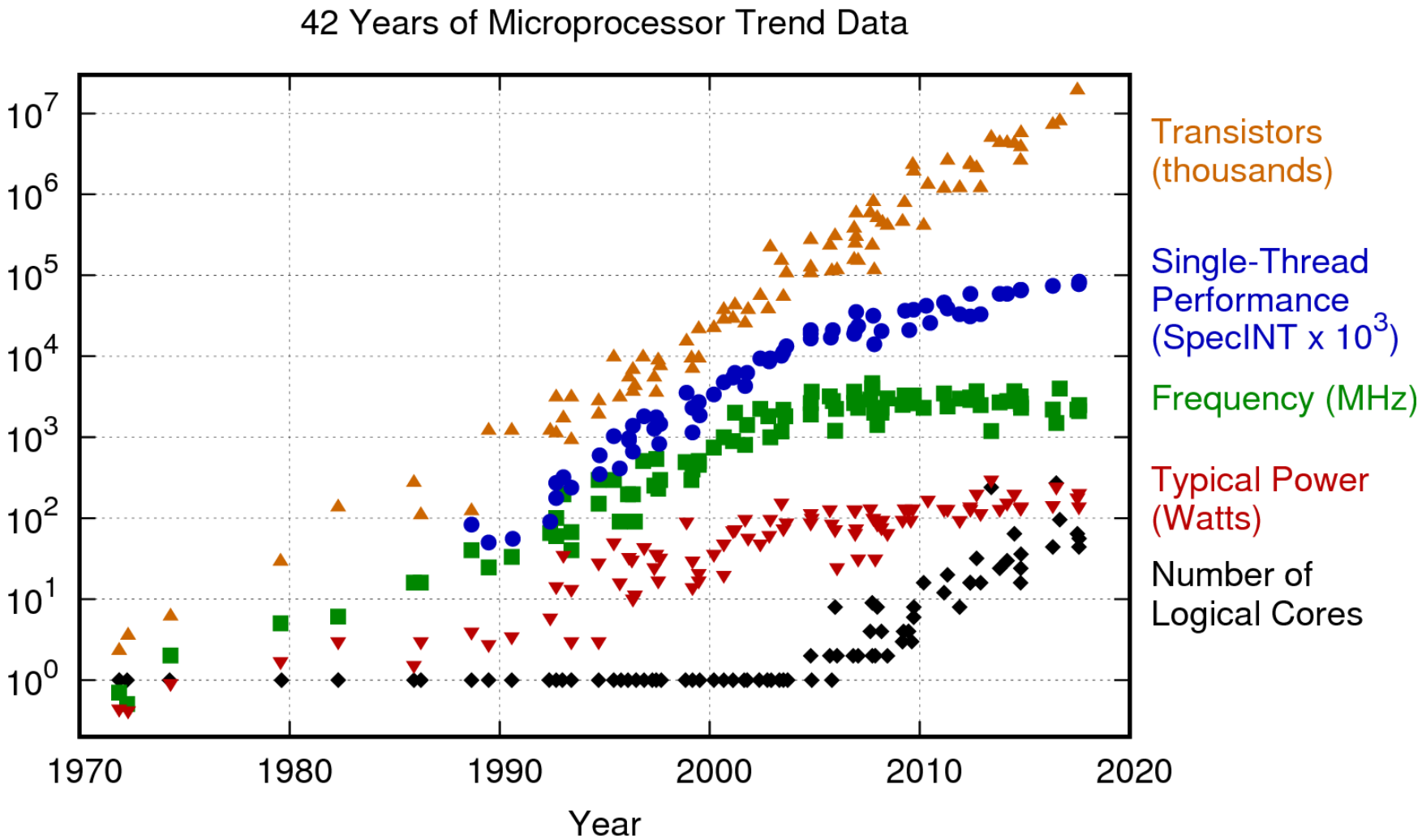
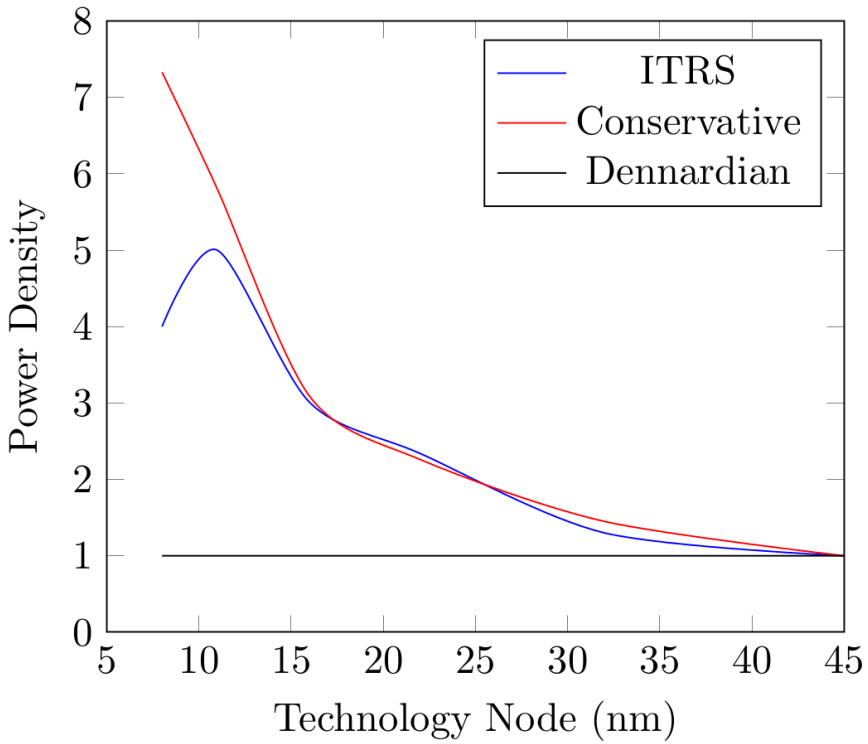


Image source: https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/

The end of Dennardian scaling

Dennardian scaling law: chip power density (power consumption per unit of area) theoretically stays constant when scaling down transistor size and voltage.



International Technology Roadmap for Semiconductors (2013)[1] and conservative Borkar projections (2010)[2] present an exponential rise in power density.

Image source: Kanduri, Anil, et al. "A perspective on dark silicon." In: *The Dark Side of Silicon*. Springer, Cham, 2017. p. 3-20.

[1] Wilson, Linda. International technology roadmap for semiconductors (ITRS). *Semiconductor Industry Association*, 2013.

[2] Borkar, Shekhar. "The exascale challenge." In: *VLSI Design Automation and Test (VLSI-DAT), 2010 International Symposium on*. IEEE, 2010. p. 2-3.

Motivation and research question

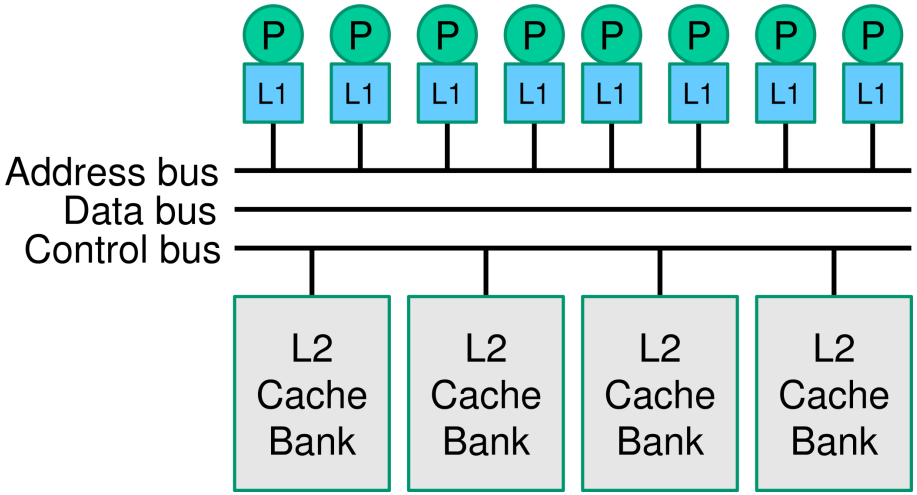
Motivation:

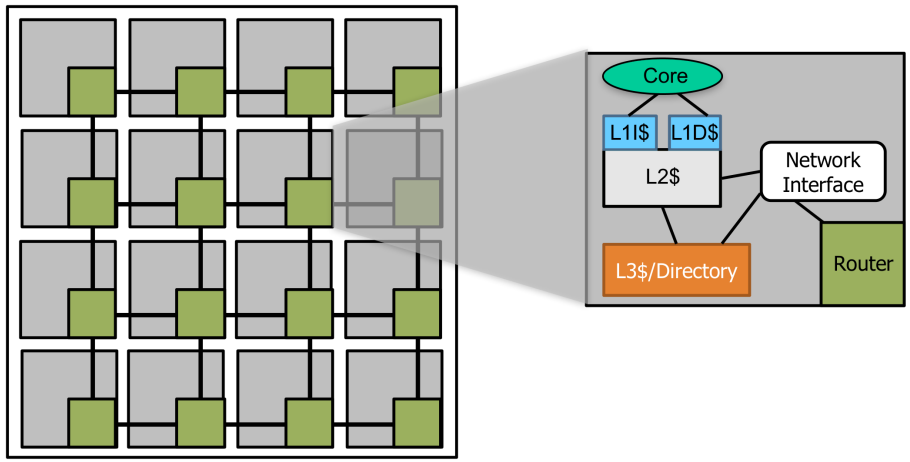
- Data must be moved between cores efficiently
- Energy-efficient and locality-optimized interconnect networks are paramount
- The classic bus interconnect scales badly
- Computer architecture simulators useful for interconnect scalability evaluation

Research question:

- To which extent do performance, power and area of various interconnect models scale with core count?

Bus versus Network-on-Chip (NoC)





Buses (shared medium)	**Networks-on-chip** (switched point-to-point)
− limited to one sender at a time	+ multiple simultaneous senders
− bandwidth is limited, shared	+ aggregate bandwidth scales with network size
− no concurrency	+ concurrent spatial reuse
− coupled computation and communication	+ decoupled computation and communication
− central arbitration	+ distributed arbitration
+ latency guaranteed if bus is granted	− routers incur extra delay
+ straightforward architecture	− complex architecture

Bus image source: Balasubramonian, Rajeev; Pinkston, Timothy M. "Buses and Crossbars." In: *Encyclopedia of Parallel Computing*. Springer US, 2011. p. 200-205.
NoC image source: `http://www.gem5.org/wiki/images/d/d4/Summit2017_garnet2.0_tutorial.pdf`

Cost scalability

Interconnect	Power dissipation	Total area	Operating frequency
$n \times n$ **mesh NoC**	$\mathcal{O}(n)$	$\mathcal{O}(n)$	$\mathcal{O}(1)$
Non-segmented bus	$\mathcal{O}(n\sqrt{n})$	$\mathcal{O}(n^3\sqrt{n})$	$\mathcal{O}\left(\frac{1}{n^2}\right)$
Segmented bus	$\mathcal{O}(n\sqrt{n})$	$\mathcal{O}(n^2\sqrt{n})$	$\mathcal{O}\left(\frac{1}{n}\right)$
Point-to-point	$\mathcal{O}(n\sqrt{n})$	$\mathcal{O}(n^2\sqrt{n})$	$\mathcal{O}\left(\frac{1}{n}\right)$

Source: Bolotin, Evgeny, et al. "Cost considerations in network on chip." *INTEGRATION, the VLSI journal*, 2004, 38.1: 19-42.

Employed computer architecture simulators

Simulators used for evaluation of interconnect scalability:

- Sniper x86 simulator v6.1
  - Package+interconnect power and area modeling by McPAT v1.3
- gem5/Garnet2.0 NoC simulator (April 27, 2018)
  - NoC power and area modeling by DSENT v0.91

Built extensions

Sniper:

- Configuration files for Haswell and Knights Landing architectures

gem5/Garnet2.0:

- Several framework assist scripts
- Topology visualization with LaTeX/TikZ
- Topologies added: line, fully connected, ring, hierarchical ring, flattened butterfly
- Configurable concentration factor $C{=}n$
  $\Rightarrow$ each router connects to n CPUs

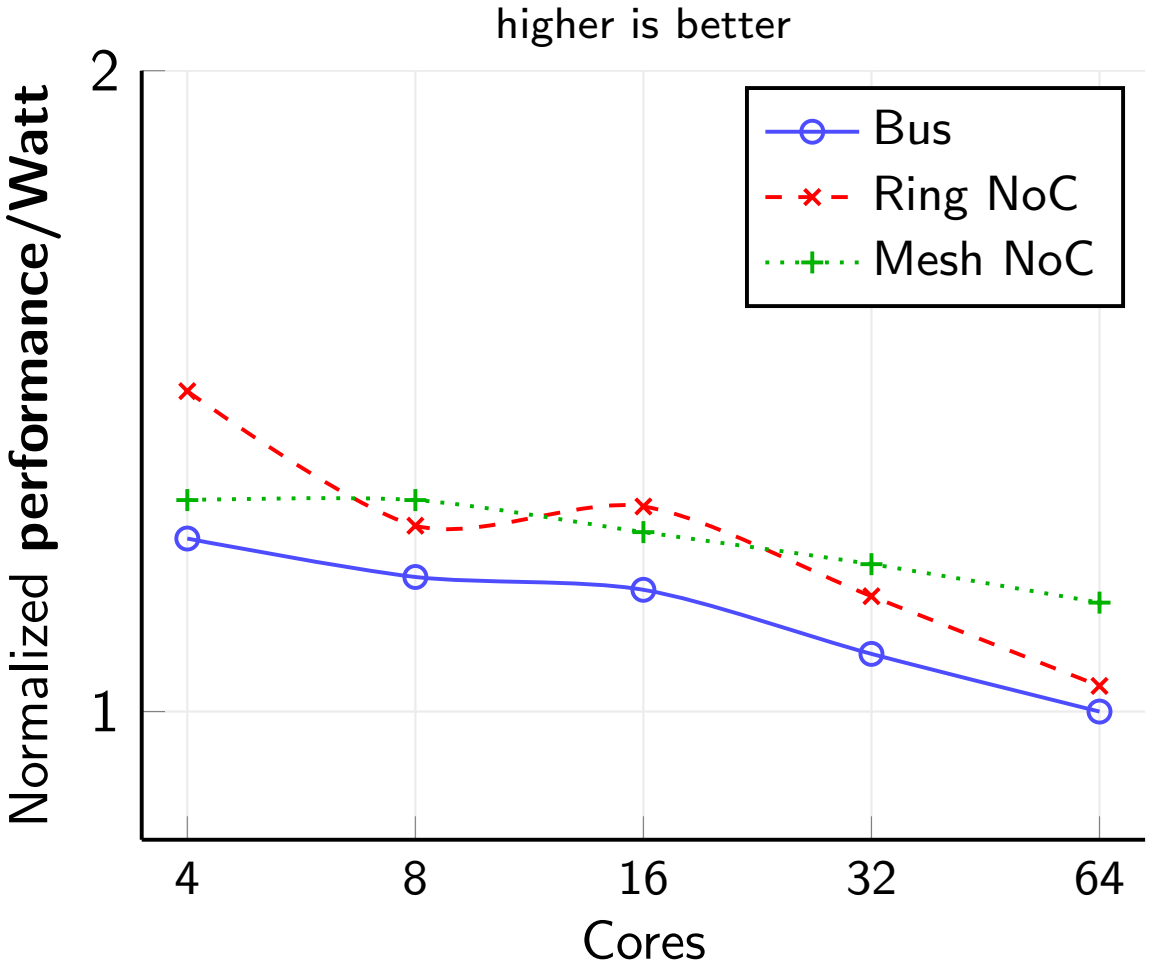
DSENT:

- Integration with Garnet2.0
- Simplified die area model

Source code freely available at: `https://github.com/Davxx/gem5_Garnet2.0_extensions`

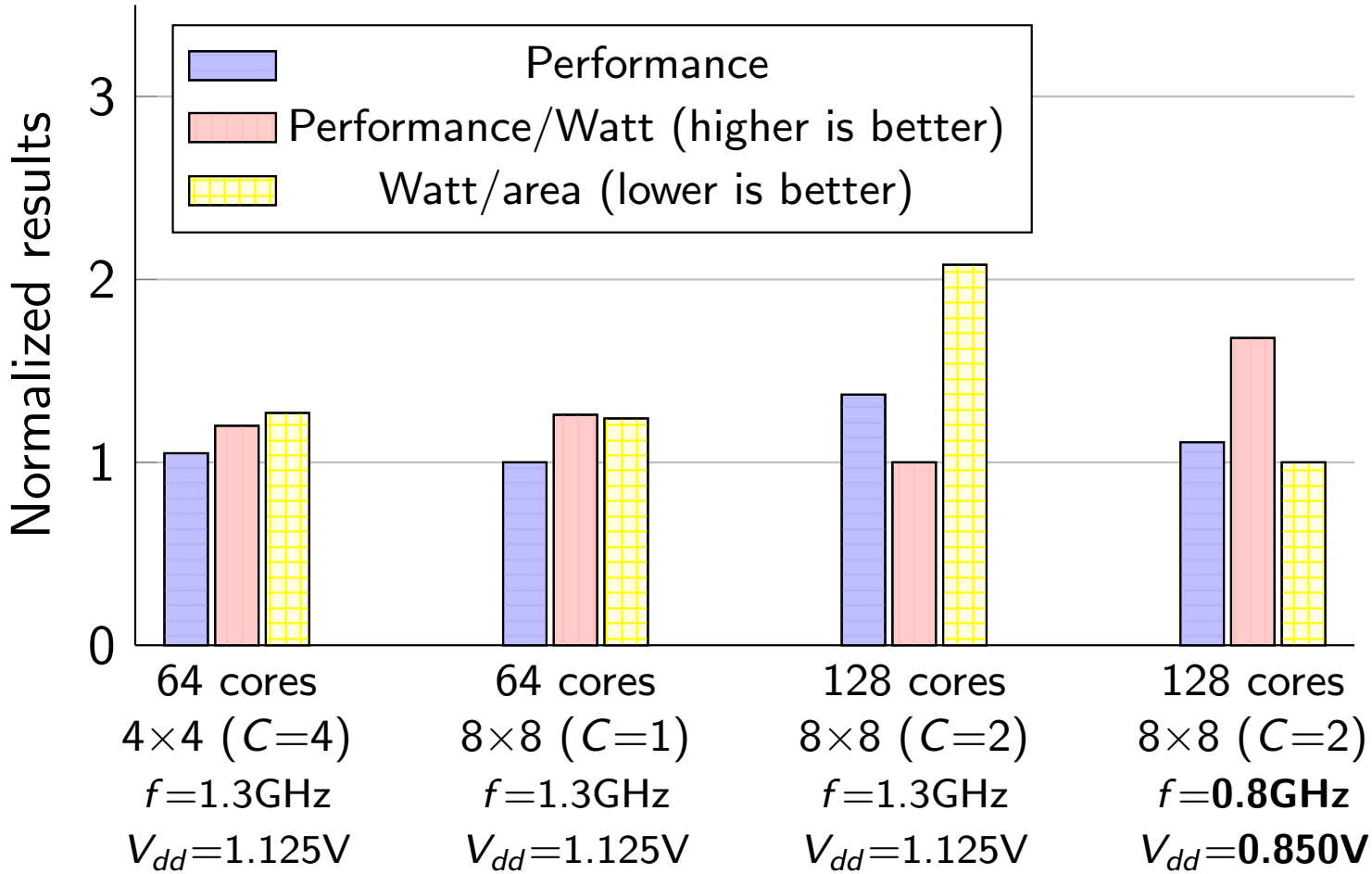
Sniper/McPAT: Haswell interconnect scalability

Sniper/McPAT results for the x86 Splash2 FFT benchmark for simulated 22 nm Haswell architectures.

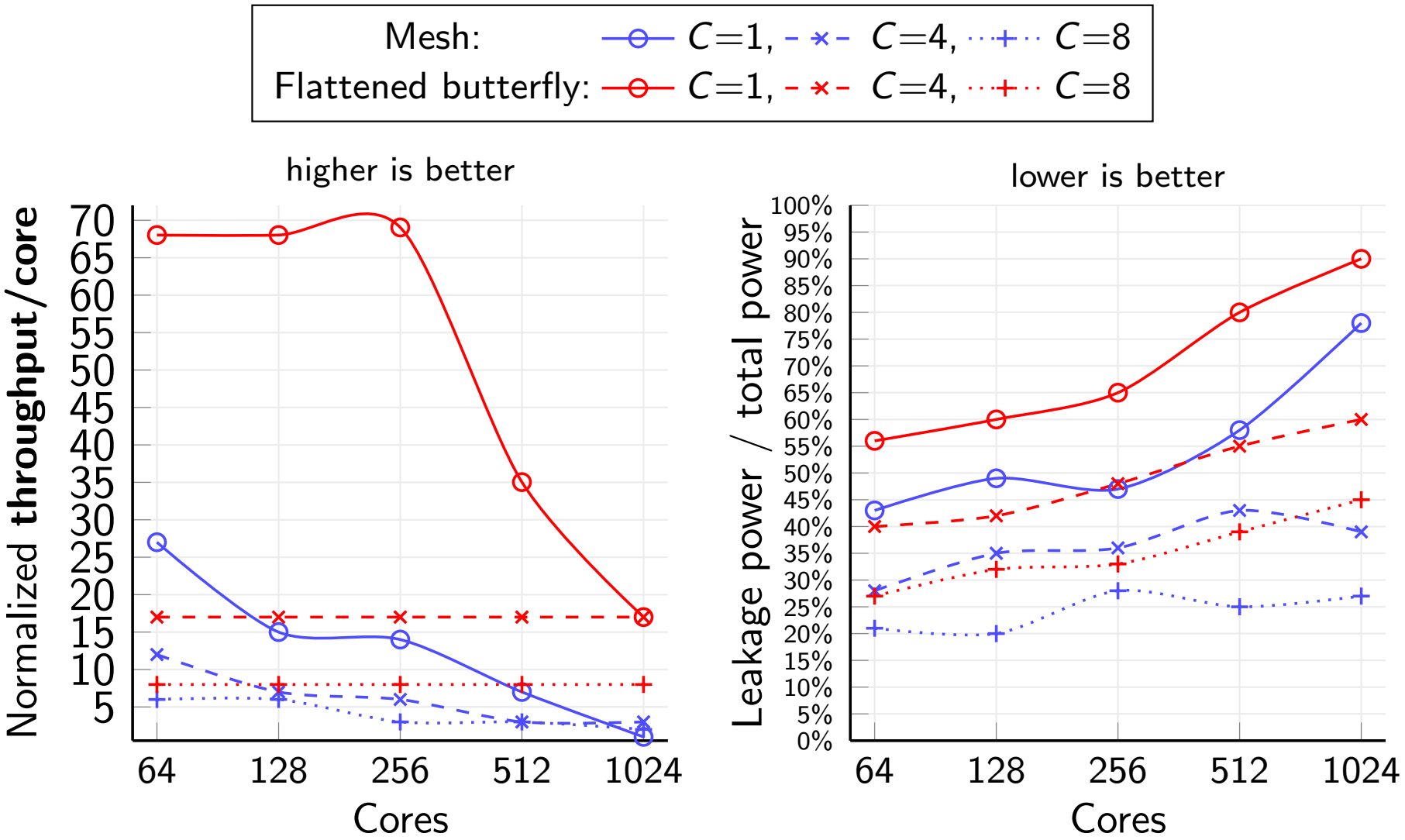


Sniper/McPAT: Knights Landing many-core scalability

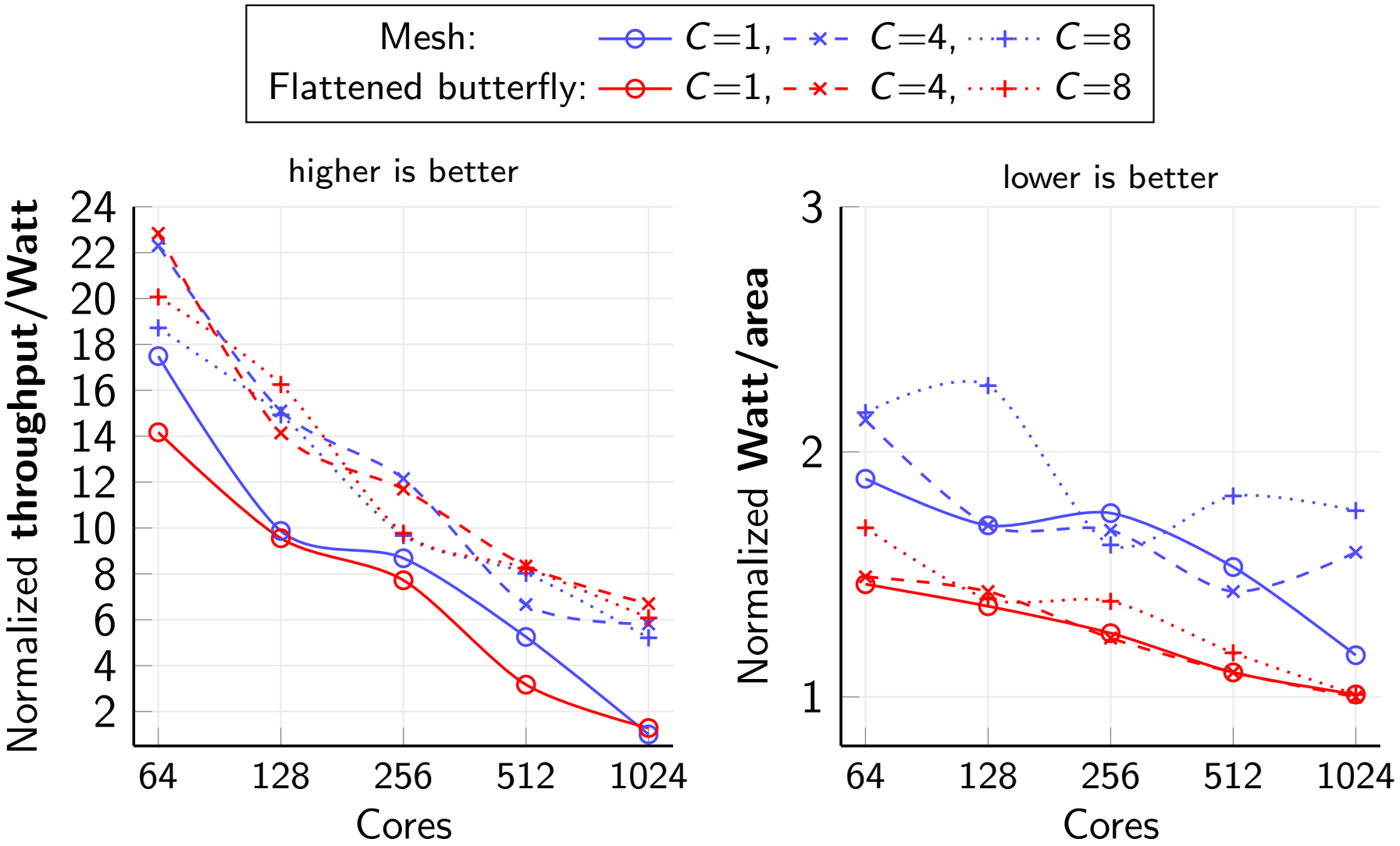
Sniper/McPAT results for the x86 Splash2 FFT benchmark for simulated 22 nm Knights Landing architectures. C=4 denotes 4 CPUs per router.



Garnet2.0: near-saturation synthetic uniform random traffic for 64-1024 cores (1 of 2)



Garnet2.0: near-saturation synthetic uniform random traffic for 64-1024 cores (2 of 2)



Conclusions: theory

- Power limits induce continuous growth in CPU core count and dark silicon
- Classic bus interconnect insufficiently scalable[3]
- Proven scalable, NoC presents a viable interconnect for the imminent many-core era
  - Researching NoC energy reduction techniques is paramount

[3] Udipi et al. propose a promising hierarchical bus-based interconnect model for up to 64 cores in: "Towards scalable, energy-efficient, bus-based on-chip networks." In: *High Performance Computer Architecture (HPCA), 2010 IEEE 16th International Symposium on*. IEEE, 2010. p. 1-12.

Conclusions: Sniper experiments

- Sniper/McPAT presently inadequate for interconnect scalability evaluation
- 128-core Knights Landing experiments show that decreasing frequency and voltage can effectuate a 1.7× gain in performance/Watt and a 2.1× gain in Watt/area, at the cost of a 19% loss in performance.
- Near-saturation synthetic traffic useful in comparing interconnect topologies

Conclusions: Garnet2.0 experiments

- Concentrated mesh and flattened butterfly are viable many-core topologies
- For 64, 128, 256, 512 and 1024 cores, at concentration factors of 4 and 8, flattened butterfly achieves average **maximum sustainable throughputs** 1.4×, 1.9×, 2.8×, 3.8× and 4.8× greater than the mesh, at an increase in **NoC area** of 1.8×, 2.5×, 3.5×, 4.0× and 6.9×, respectively
- Flattened butterfly's higher throughputs incur more dynamic power consumption; its large amount of buffers and links incur more leakage power
- Flattened butterfly still achieves superior scalability, marked by its substantially superior Watt/area
  - Higher cost can be a decisive factor in foregoing the flattened butterfy in favor of less complex topologies

Appendix I: Sniper simulation parameters

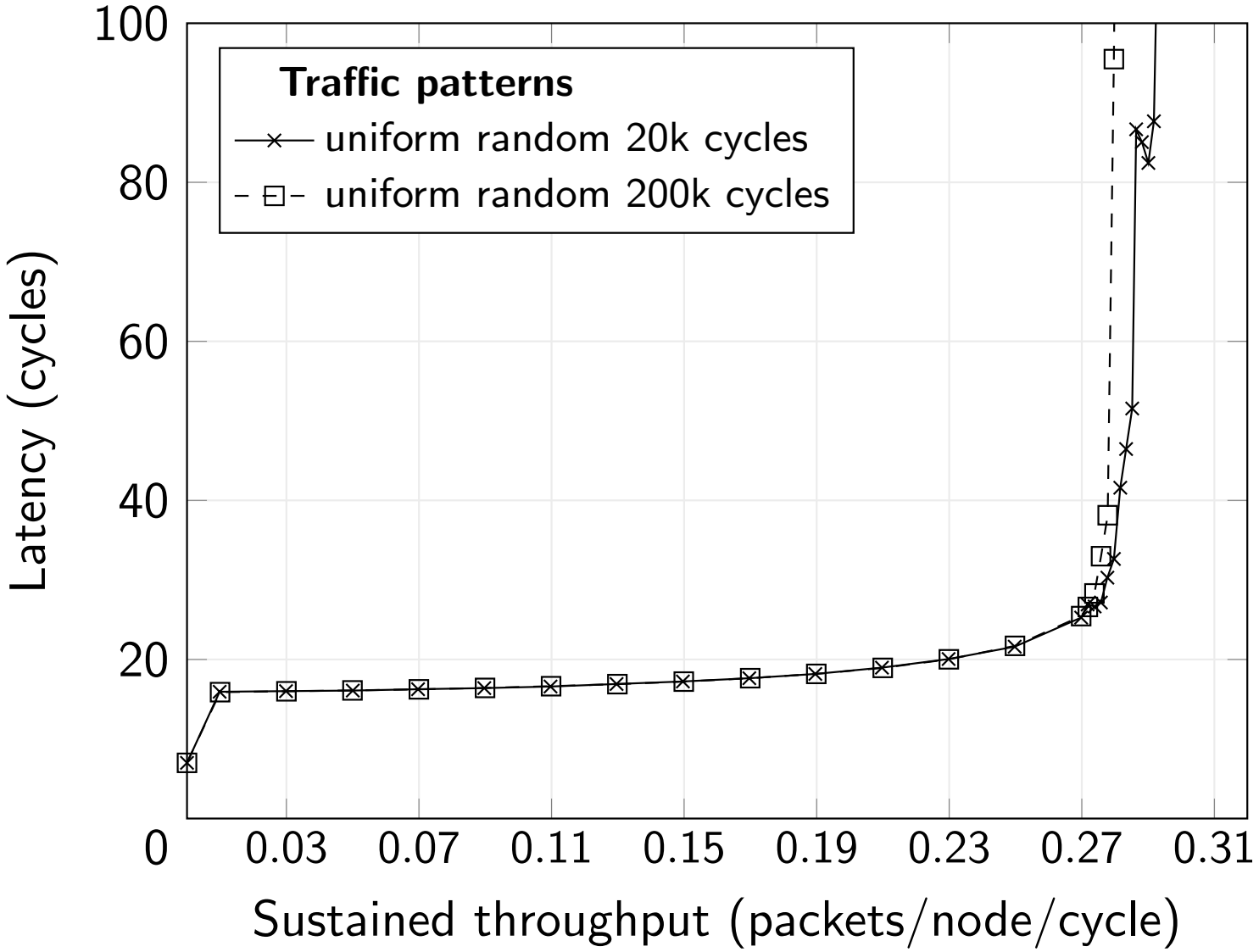
Knights Landing architecture:

- out-of-order ROB core model
- 2 DRAM controllers $\times$ 51 GB/s bandwidth
- 8-way associative 32 KB private L1d and L1i caches
- 16-way associative 1 MB shared L2 cache
- n$\times$n mesh, 512-bit link bandwidth
- MESIF cache coherency protocol

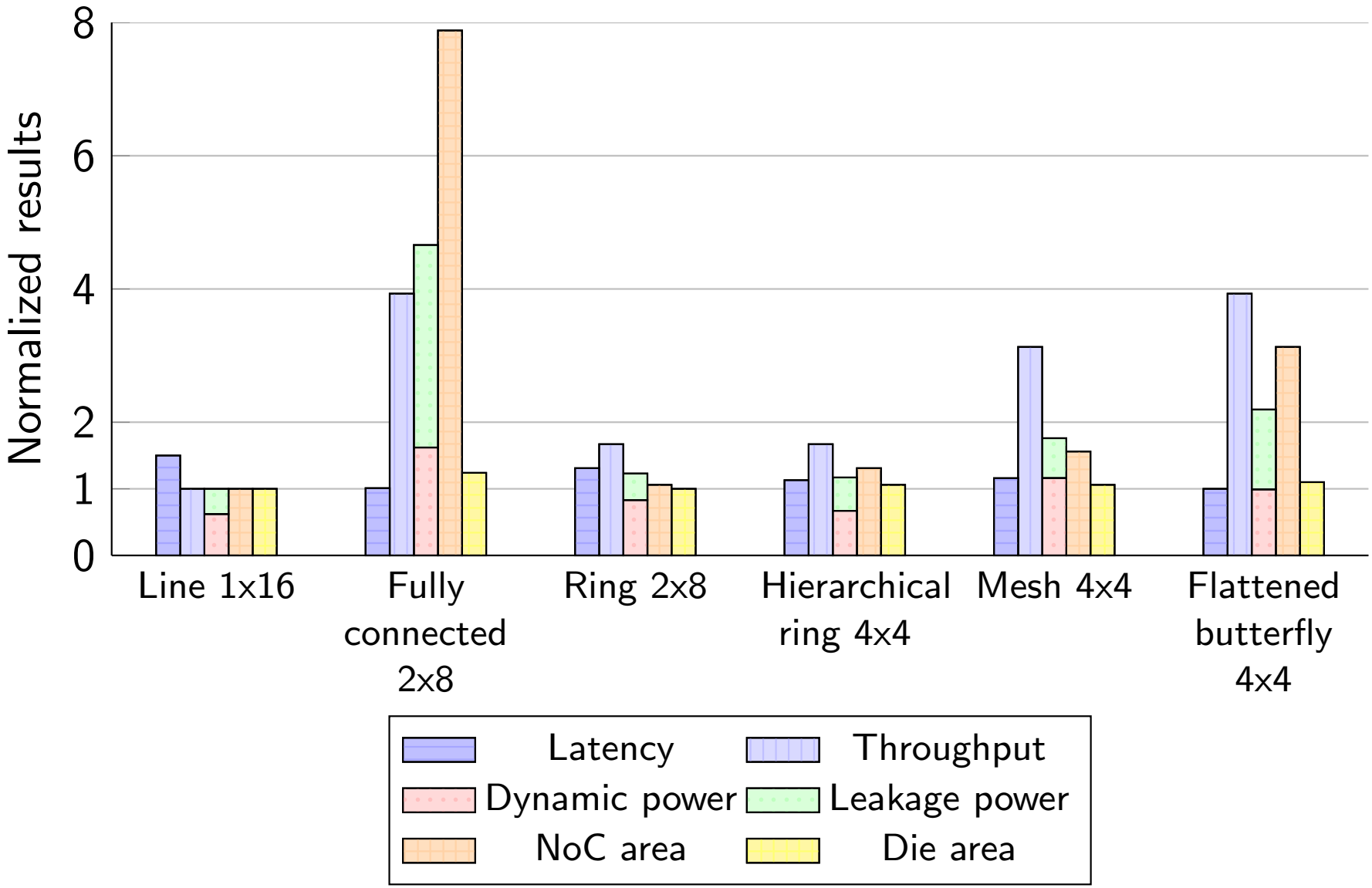
Appendix II: Garnet2.0 simulation parameters

- 4 VCs per VNET
- 1 buffer per control VC, 8 buffers per data VC
- 1-flit sized control packets (VNETS 0 and 1) and 5-flit sized data packets (VNET 2)
- 512-bit link width
- uniform random injection into all three VNETs
- injection rate: topology-specific near-saturation
- each CPU is connected through one L1 cache controller node
- the number of directory nodes equals the number of routers, with a maximum of 256
- DSENT technology model: 11 nm tri-gate

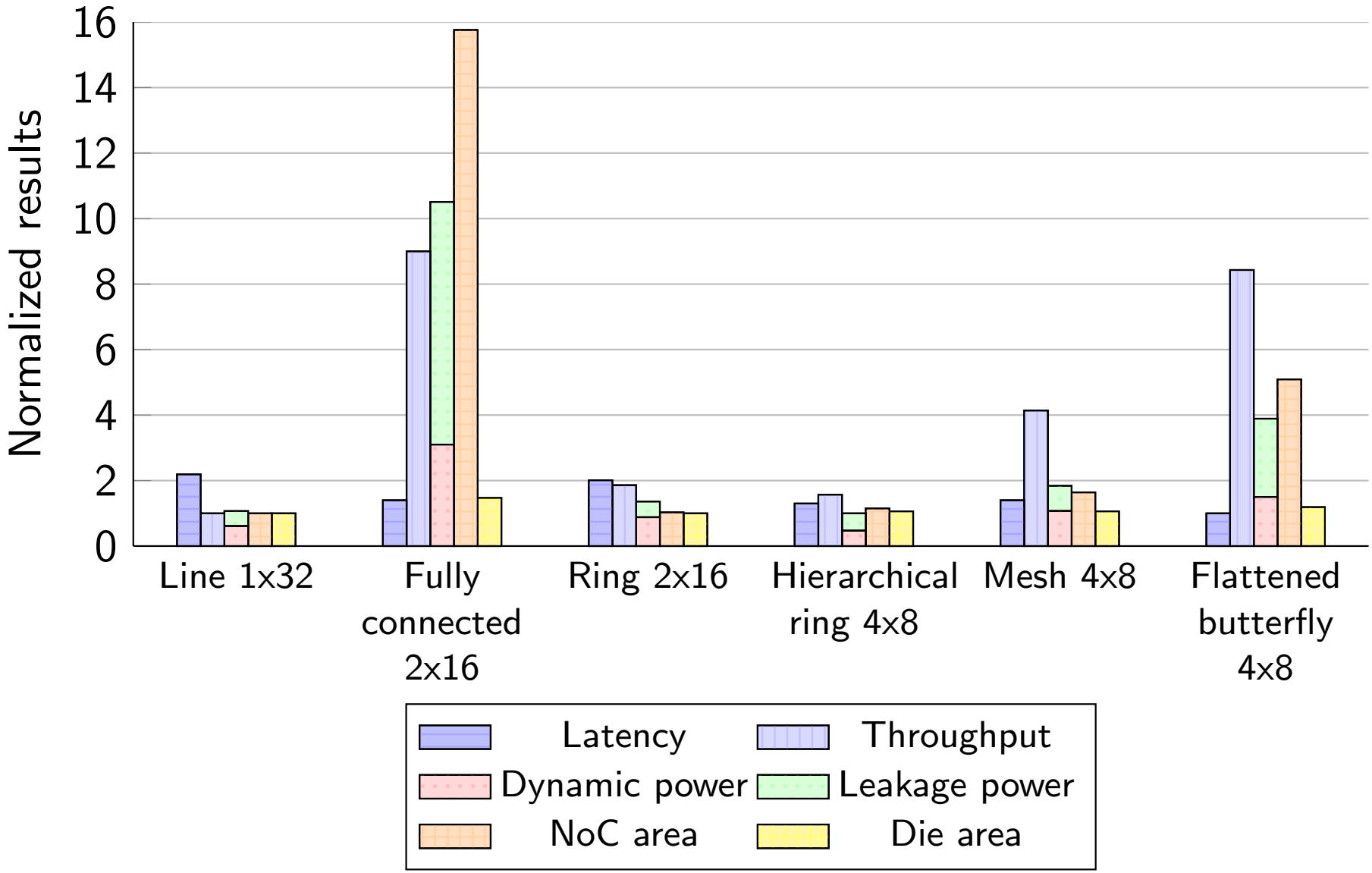
Appendix III: Garnet2.0 network saturation



Appendix IV: Garnet2.0 results for 16 core topologies



Appendix V: Garnet2.0 results for 32 core topologies



Appendix VI: Garnet2.0 TikZ visualization example

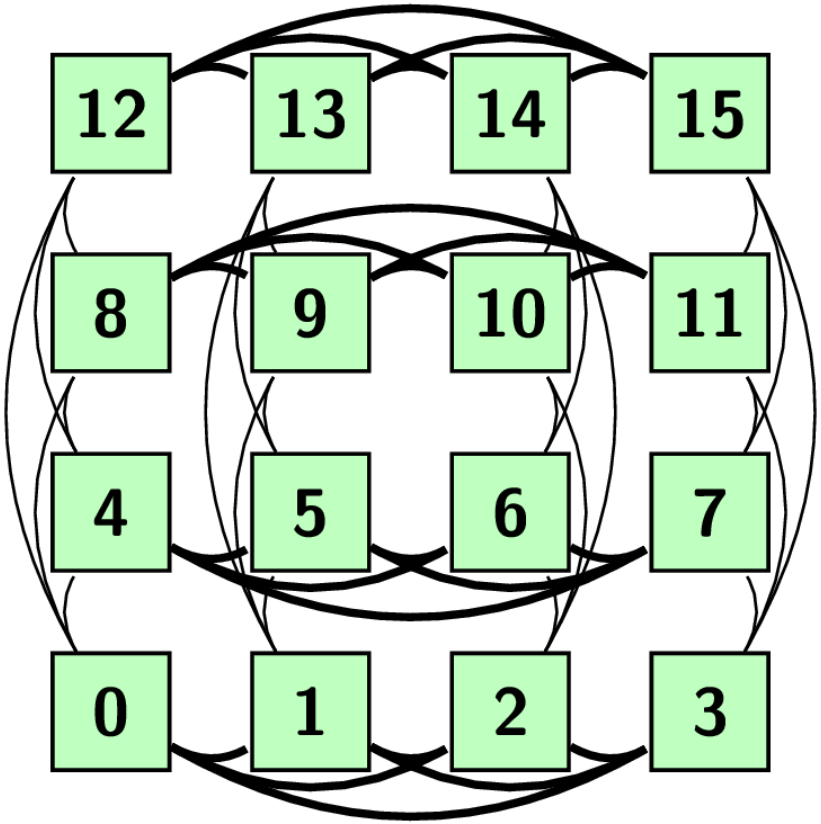
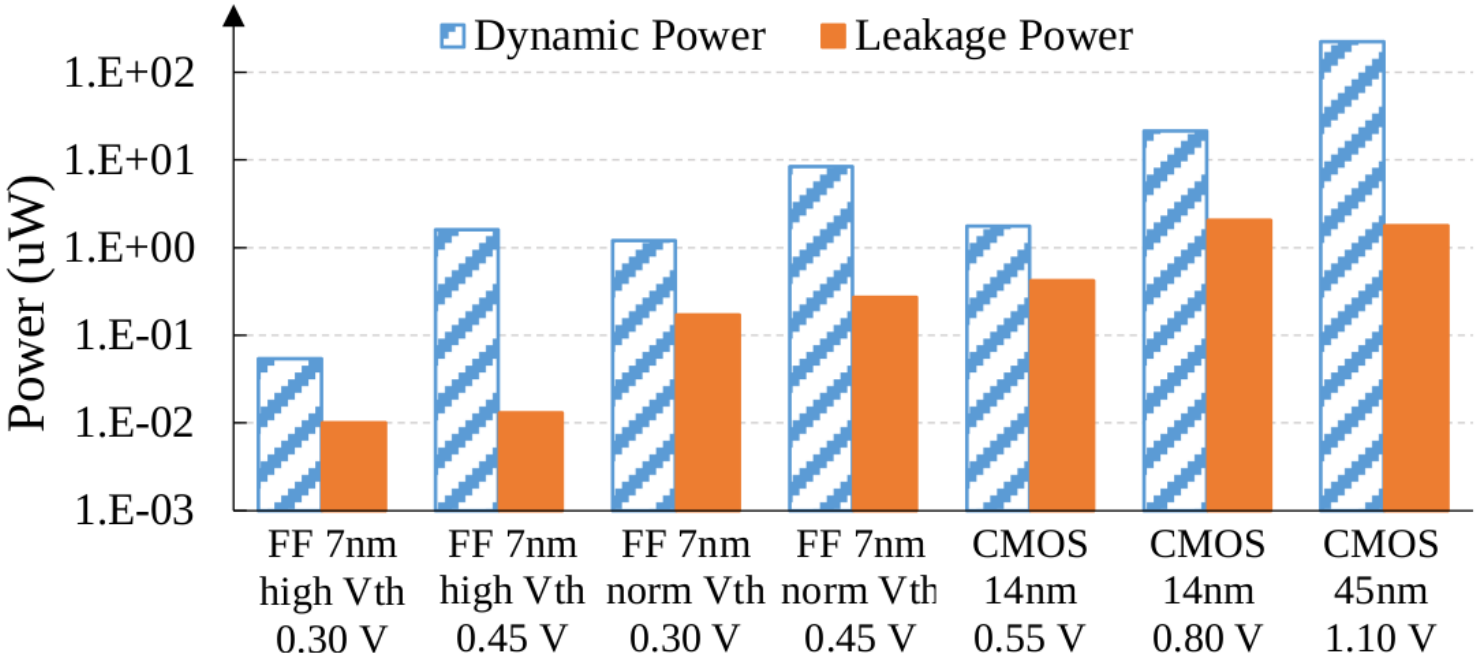


Diagram of the 16-router flattened butterfly topology, generated by the newly created `TikzTopology` class. Each router can be connected to n CPUs, where n is the concentration factor. The flattened butterfly debut paper (2007) suggests concentration factors of $n = 4$ for 64 cores and $n = 8$ for 128 cores.

Appendix VII: FinFET versus CMOS dynamic and leakage power consumption



Dynamic and leakage power consumptions in the c432 benchmark circuit for proposed 7 nm FinFET standard cell libraries and conventional bulk CMOS standard cell libraries.

Image source: XIE, Qing, et al. "Performance comparisons between 7-nm FinFET and conventional bulk CMOS standard cell libraries." *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2015, 62.8: 761-765.