SAND2019-0427PE

HBM2/3 Evaluation on Many-Core CPUs

Memory Working Group
Hardware Evaluation

Gwen Voskuilen[1], Alfredo Gimenez[2], Ivy Peng[3], Shirley Moore[3], Maya Gokhale[2]
[1]SNL, [2]LLNL, [3]ORNL

ECP Annual Meeting
1/15/2019

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Exascale Memory Systems

- Memory architecture will key role in delivered performance of exascale machines
  - Organization, bandwidth, latency, capacity, power affect usability and capability
  - High cost and power usage of DRAM require optimizing packaging and organization
- High-bandwidth memory (HBM) will feature prominently in exascale machines
  - Higher bandwidth than conventional DDR at similar latencies
- Next generation HBM3 is expected in the ~2020 time frame

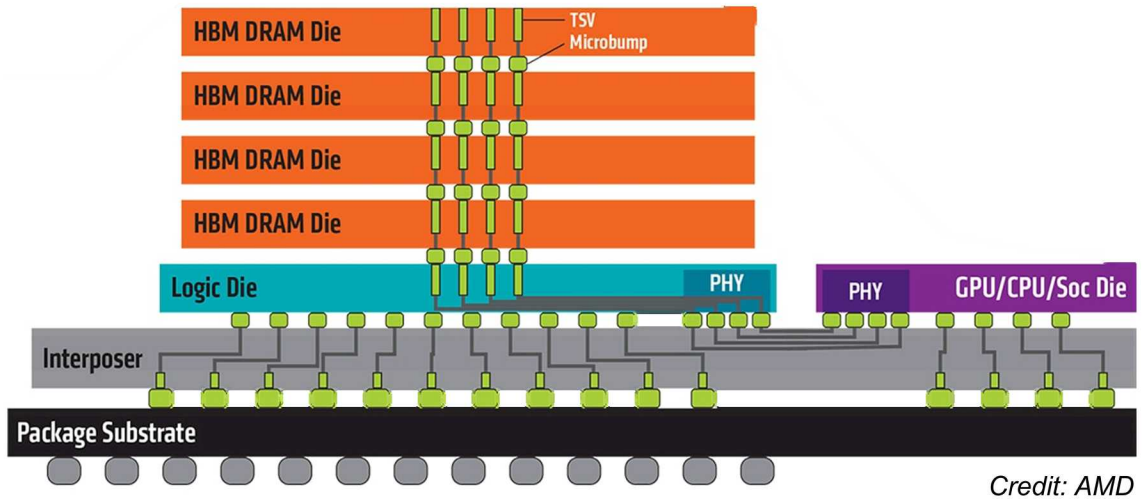
Key question:

To what degree will HBM3 benefit application performance beyond HBM2?

EXASCALE COMPUTING PROJECT

High Bandwidth Memory (HBM)

- HBM2
  - 256GB/s
  - Up to 8 dies/stack @ 8Gb/die
    - < =8GB / stack
- HBM3
  - Future generation, expected ~2020
    - JEDEC standard not finalized
  - 512GB/s
  - > 8 dies/stack @ 16Gb/die

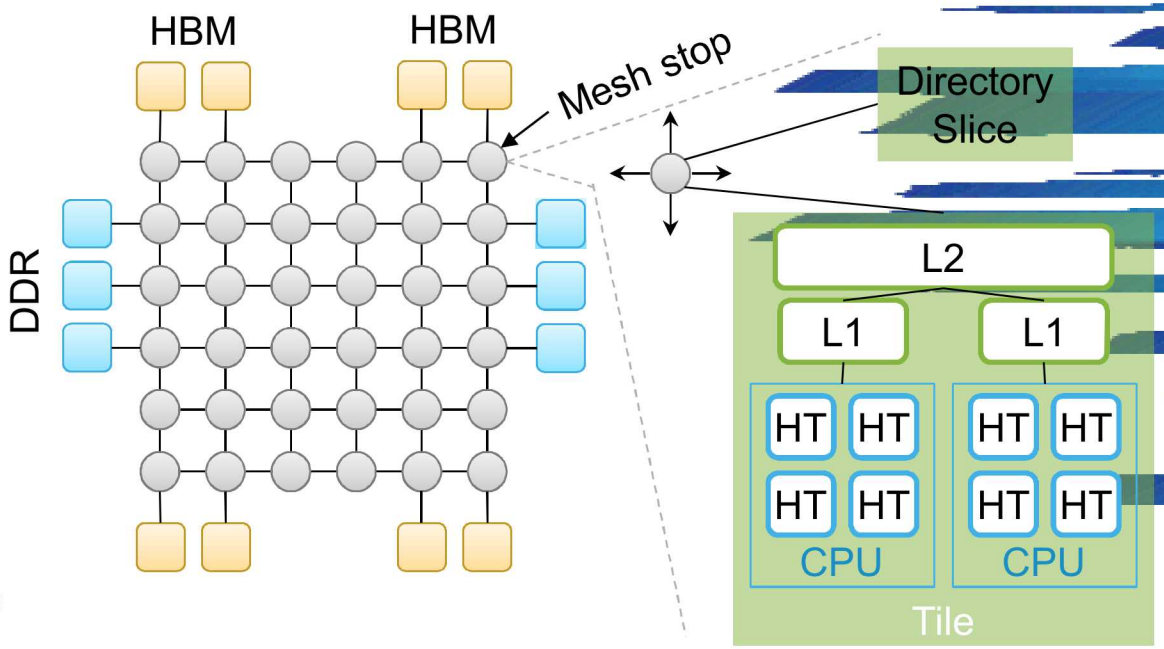


Credit: AMD

Compare performance of HBM2 and HBM3 for exascale applications

Methodology

- HBM3 hardware not available , so use simulation
- Used Structural Simulation Toolkit (SST)
  - Framework that ties together many models of different architectural features
    - E.g., cores, caches, memory, network-on-chip
- Use Knight's Landing (KNL) as a baseline
  - Enables model calibration against hardware
  - KNL's MCDRAM ≈ HBM in bandwidth/latency



Methodology: Benchmarks and simulation

- Used proxy apps to evaluate
  - Reasonable simulation target
    - Versus kernels (too small) and apps (too big)
  - Stream, HPCG, XSBench
  - Single process, 256 threads

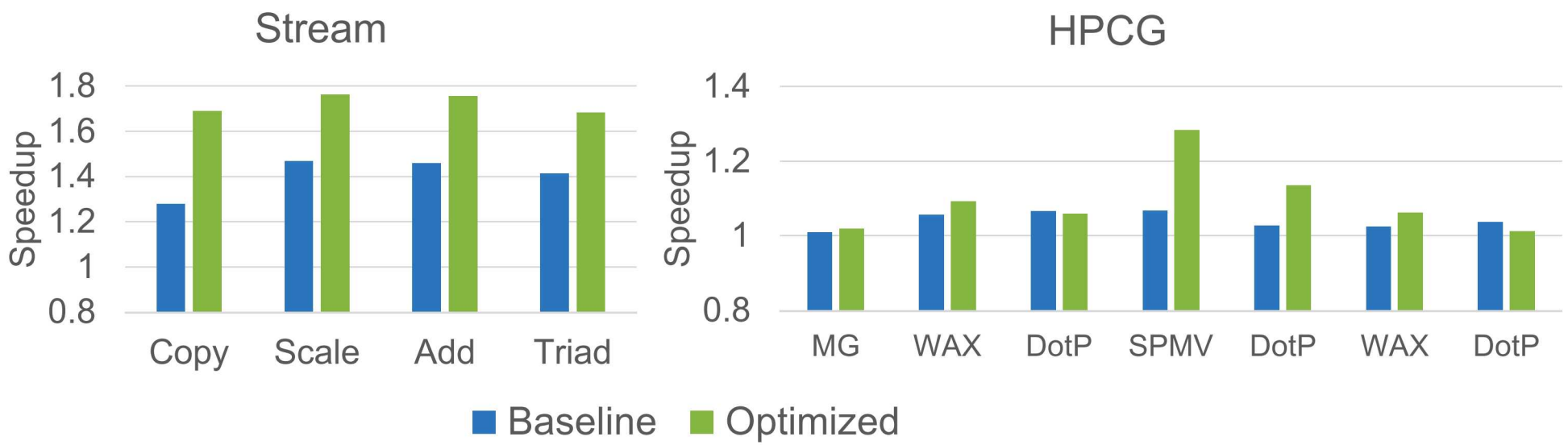
Benchmark Parameters	
Stream	10-100M elem/array, 2-10 iterations
HPCG	272x272x272, 88x88x88
XSBench	'large' with 1M lookups

- Simulated Architectures

	Calibration	HBM2	HBM3	HBM2-opt	HBM3-opt
HBM	256 GB/s	256 GB/s	512 GB/s	256 GB/s	512 GB/s
Network-on-chip	*1.7 GHz (KNL)*	3.4 GHz	5.1 GHz	6.8 GHz	6.8 GHz
Throughput limits	*KNL*	*KNL*	*KNL*	Unlimited	Unlimited
Request volume limits	*KNL*	*KNL*	*KNL*	1.3X+	1.3X+

EXASCALE COMPUTING PROJECT

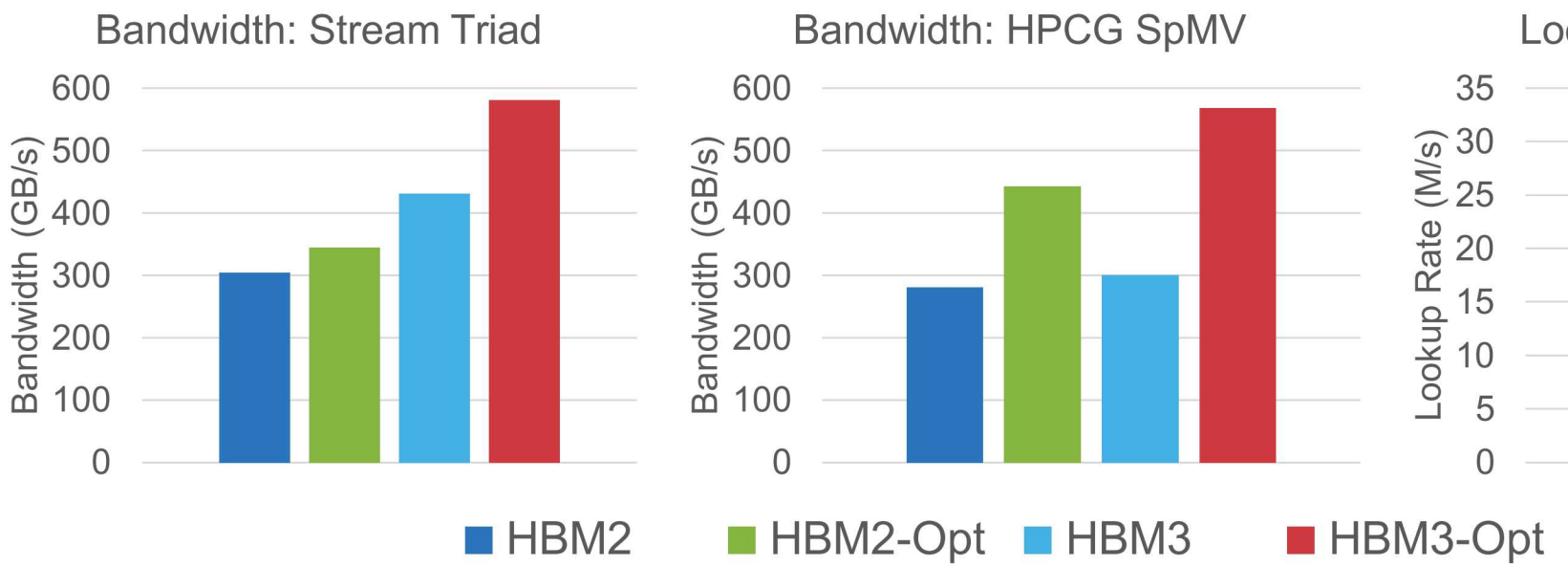
Results: Execution time speedup of HBM3 normalized to HBM2



- HBM3 alone (blue) yields small to moderate performance gains
- Architecture improvements enable HBM3 to significantly outperform HBM2 on Stream, moderately outperform on HPCG and XSBench

HBM3 requires significant core/cache-subsystem improvements to benefit applications

Results: Performance vs. Architecture and Memory



- Stream benefits from HBM3 alone, greater benefit from optimized architecture
- HPCG and XSBench do not benefit from HBM3 without optimized architecture
  - HBM2 benefits substantially from optimized architecture as well

Observations and Recommendations

- **Observations**
  - The network-on-chip (NoC) is a bandwidth bottleneck
    - KNL, simulated architectures
  - Without the NoC bandwidth limitation, HBM3 provides modest gains in the best case (Stream)
- **Recommendations**
  - Processor (core/cache/NoC) improvements should be prioritized ahead of HBM3
    - Specifically, focus on the NoC and increasing the volume of outstanding memory requests throughout the cache subsystem
  - In the absence of improvements, increased capacity will be the primary advantage of HBM3

EXASCALE COMPUTING PROJECT

Thank you!

Backup

Proxies & Benchmarks

		Configurations		
		Default	Calibration	Simulation
Stream	Array Size	100M	100M	10M
	Memory footprint	2.2GB	2.2GB	229 MB
	Iterations	10	2	2
HPCG	Input	272 x 272 x 136	272 x 272 x 136, 88 x 88 x 88	88 x 88 x 88
	Memory footprint	8 GB	8 GB, 0.5 GB	0.5 GB
	Other modification		Simulate first iteration only	
XSBench	Dataset	'large'	'large'	'large'
	Lookups	15M	1M	1M
	Memory footprint	5.3 GB	5.3 GB	5.3 GB
	Other modification		Simulate cross-section (XS) lookups only	

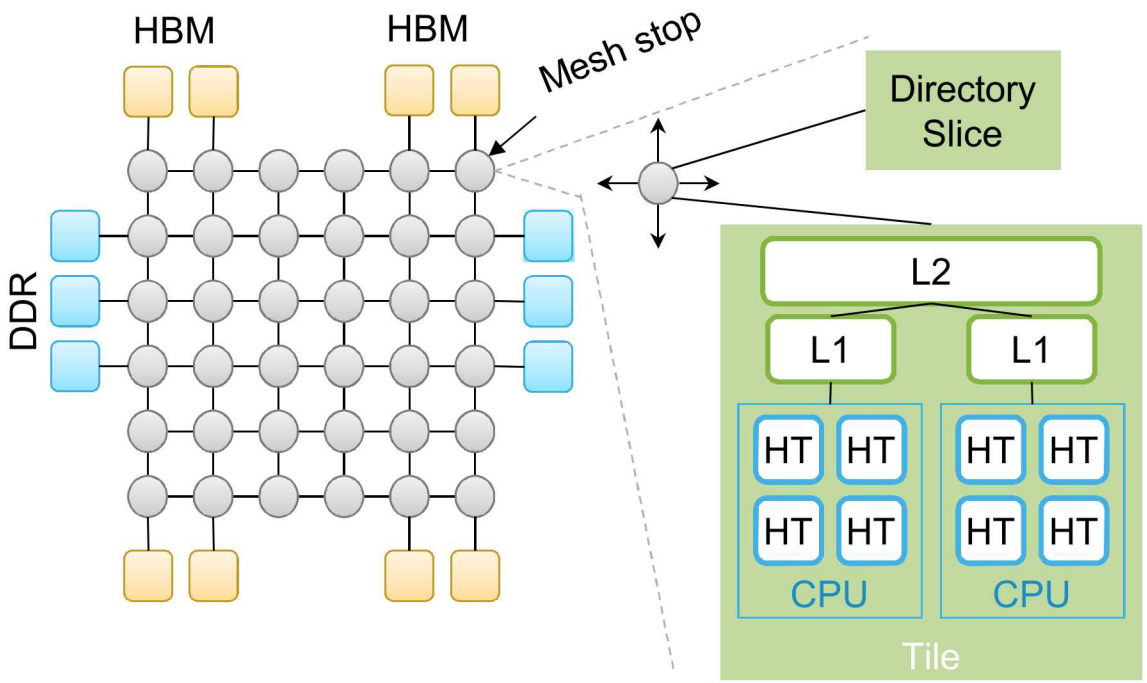
EXASCALE COMPUTING PROJECT

Simulation parameters

	Parameter	Calibration	HBM2	HBM3	HBM*-opt
CPU	Issue (instructions/cycle)	2	2	2	16
	Request Queue	12	12	12	24
L1	MSHR entries	12	12	12	24
	Requests/cycle	2	2	2	Unlimited
L2	MSHR entries	48	48	48	64
	Requests/cycle	1	1	1	Unlimited
NoC	Frequency	1.7 GHz	3.4 GHz	5.1 GHz	6.8 GHz
	Control link bandwidth	12.7 GB/s	25.3 GB/s	38 GB/s	50.7 GB/s
	Data link bandwidth	57 GB/s	114 GB/s	171 GB/s	228 GB/s
Directory	MSHR entries	128	128	128	512
HBM	Channels (2 pCh / ch)	8	8	16	*
	Bandwidth	256 GB/s	256 GB/s	512 GB/s	*

EXASCALE COMPUTING PROJECT

SST Models

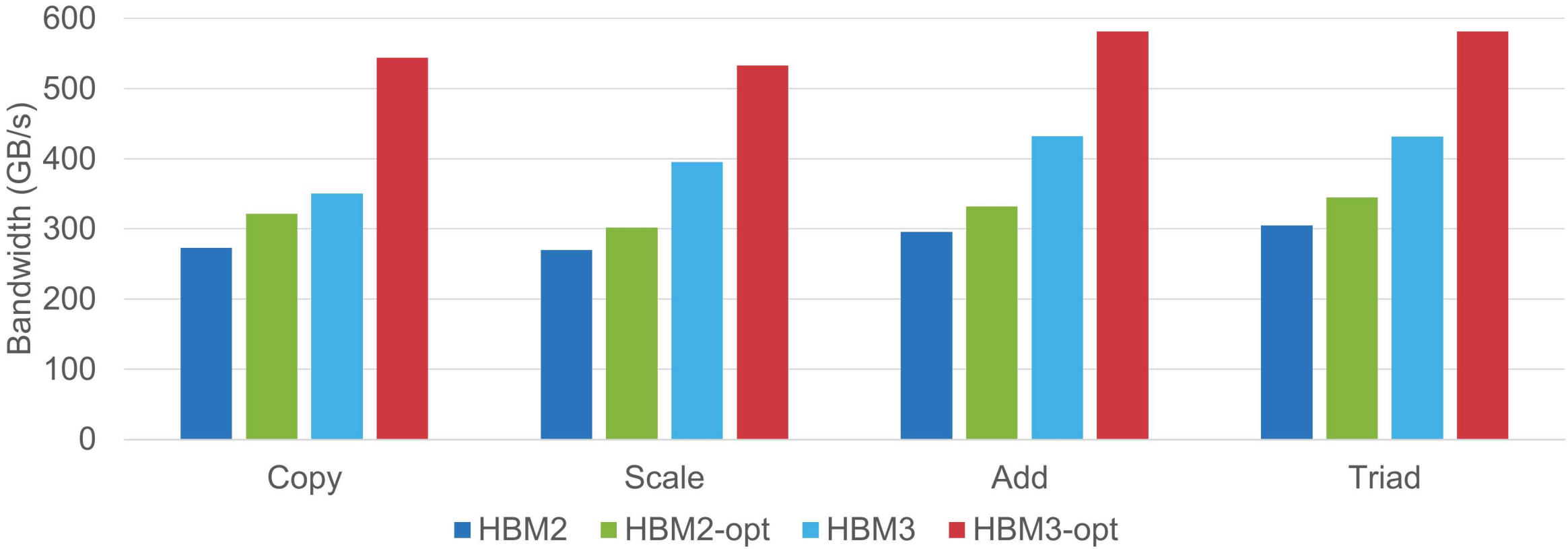


- Cores: Ariel
  - PIN based, interactive trace
- Caches: MemHierarchy
  - Prefetch: Cassini
- Directory: MemHierarchy
- NoC: Kingsley
- DDR: *Unused in this study*
- HBM: CramSim
  - HBM2 model developed by IBM

Calibration findings

- Model is accurate for bandwidth-bound codes (within ~10%) but less so for latency-bound codes
  - Models used, especially the core model (Ariel), are optimistic when it comes to concurrency
- Thus, the simulation should err on the side of overestimating HBM3 performance
  - Results can be viewed as *upper bound* on performance

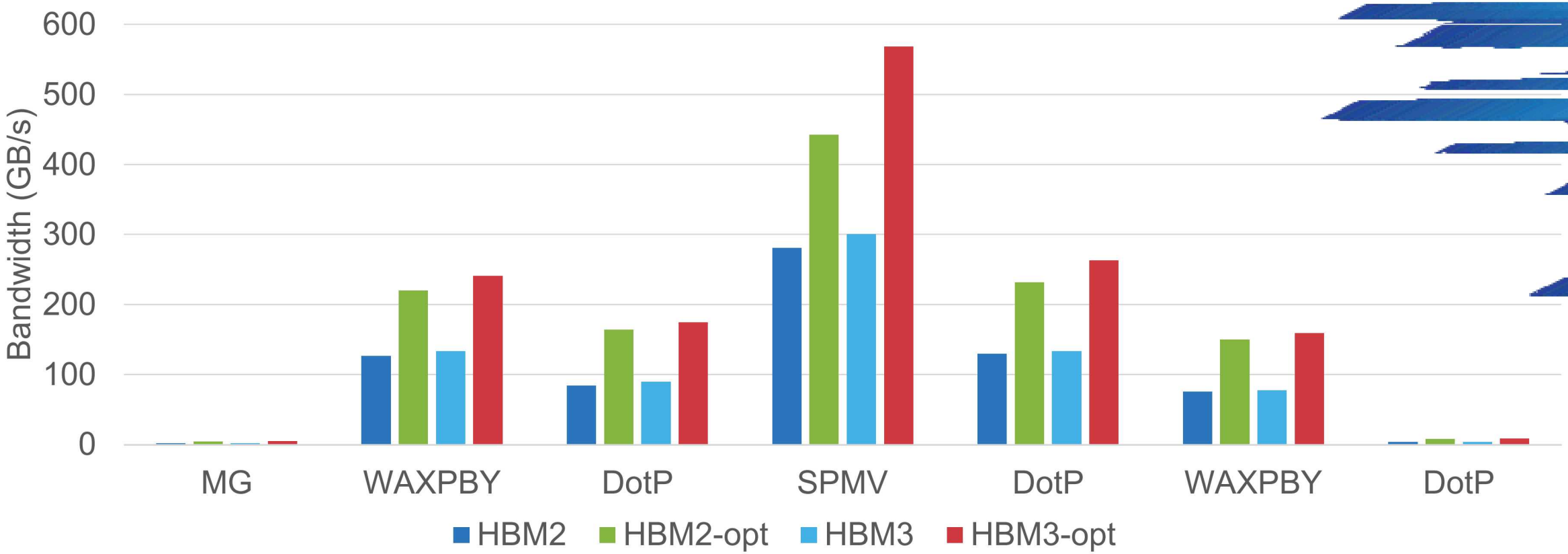
Stream Performance



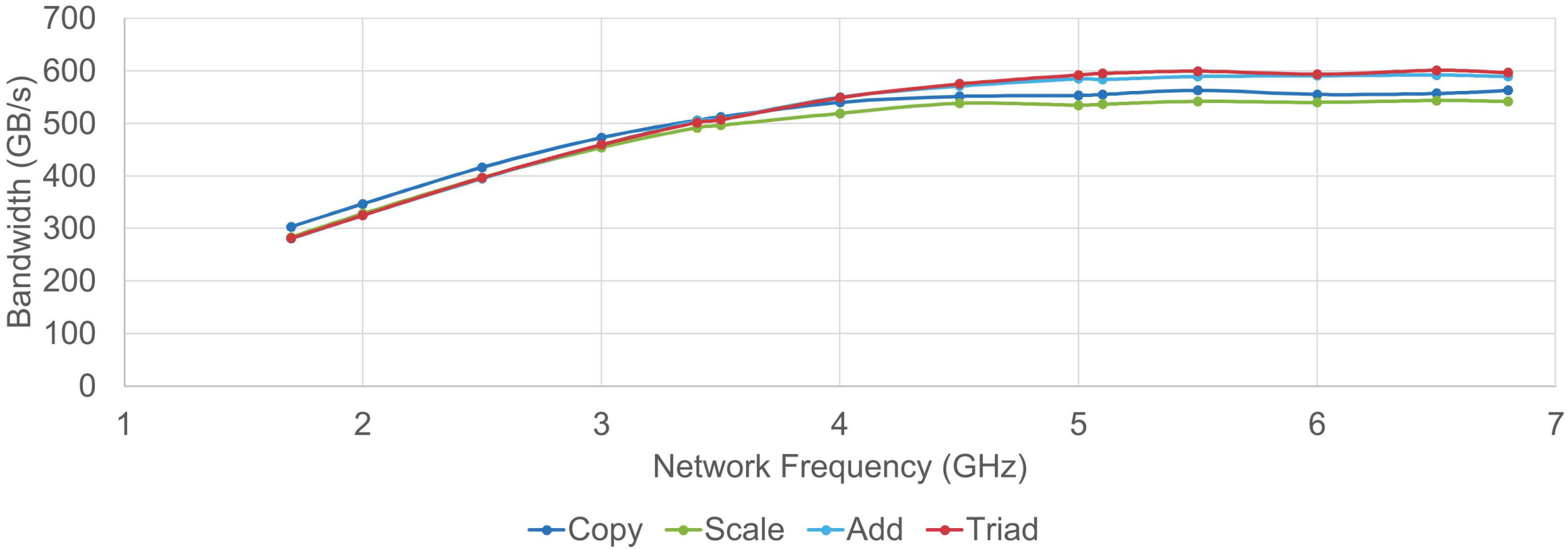
HPCG Functions

#	Function	Equation	Access Pattern
f1	ComputeMG	Gauss-Seidel Preconditioner	Irregular
f2	ComputeWAXPBY	$p = \beta * p + z$	Regular
f3	ComputeDotProduct	$rtz = r * z$	Regular
f4	ComputeSPMV	$Ap = A * p$	Irregular
f5	ComputeDotProduct	$\alpha = p * Ap$	Regular
f6	ComputeWAXPBY (x2)	$x = x + \alpha * p, r = r - \alpha * Ap$	Regular
f7	ComputeDotProduct	$normr = r * r$	Reguar

HPCG Performance

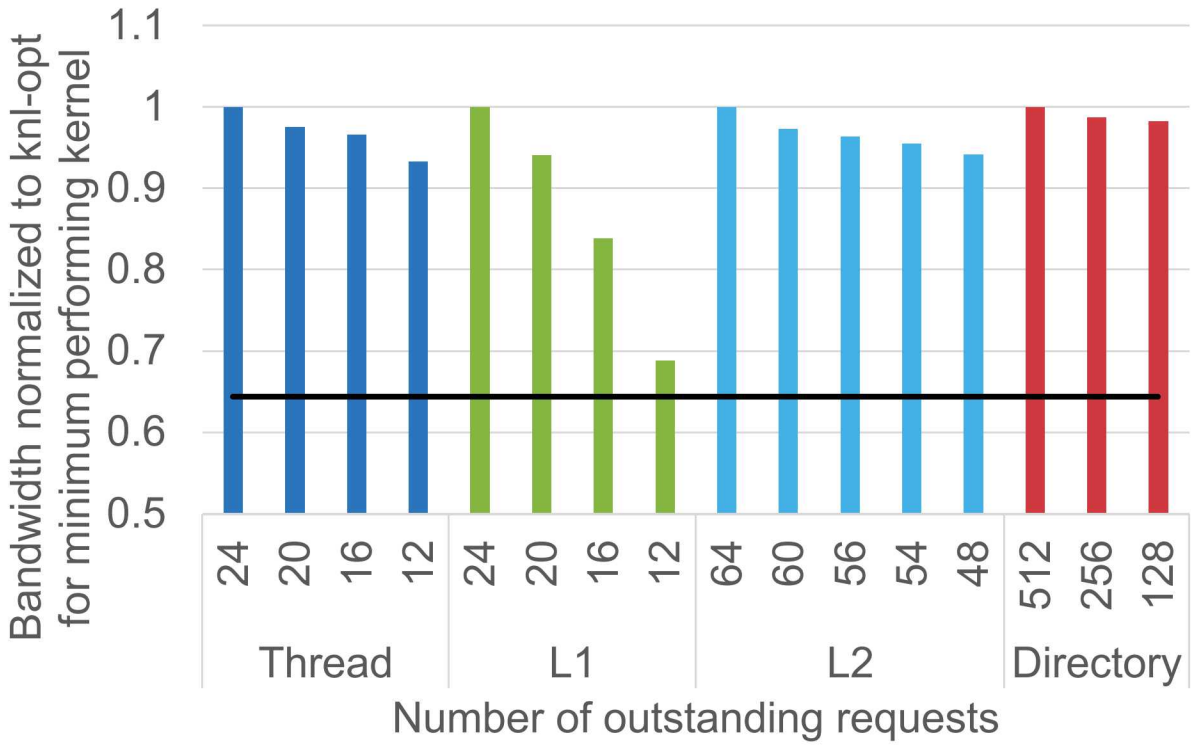


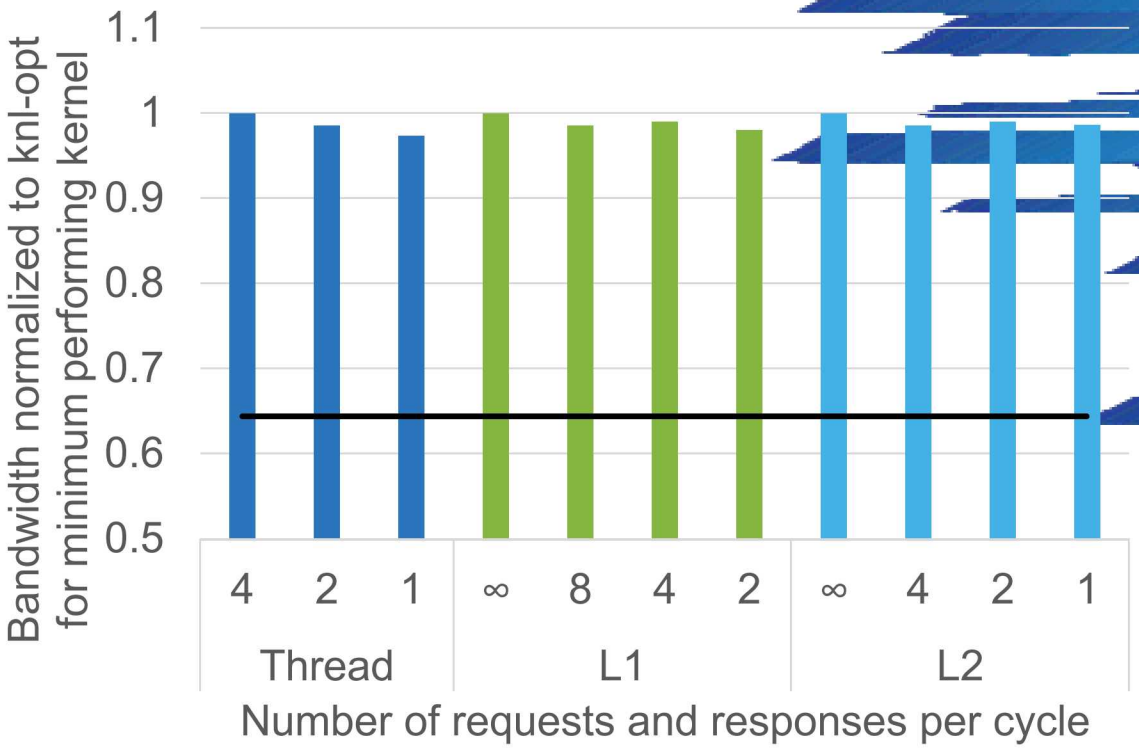
Stream Sensitivity to NoC Frequency (Bandwidth)



Using HBM3-Opt architecture

Stream sensitivity to throughput/request volume parameters





Using HBM3-Opt architecture