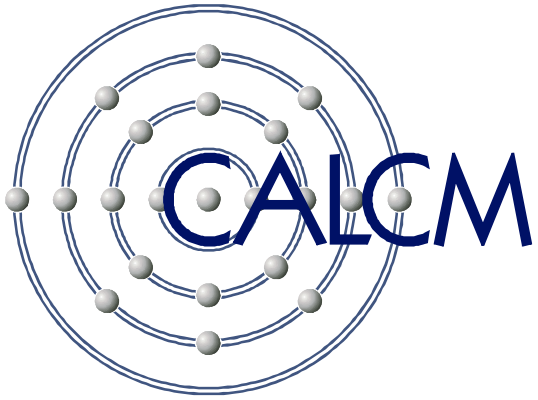
Single-Chip Heterogeneous Computing: Does the Future Include Custom Logic, FPGAs, and GPGPUs?

Eric S. Chung, Peter A. Milder,
James C. Hoe, Ken Mai

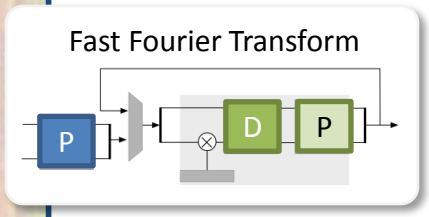


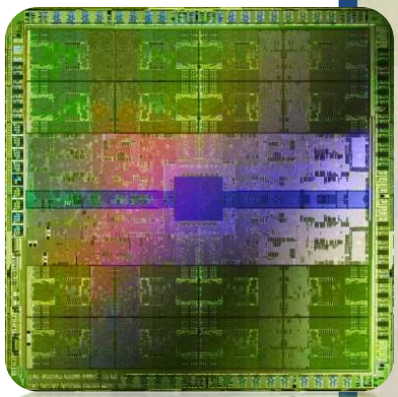
Computer Architecture Lab at
Carnegie Mellon

Will Future Multicores Become Heterogeneous?

General Purpose Core

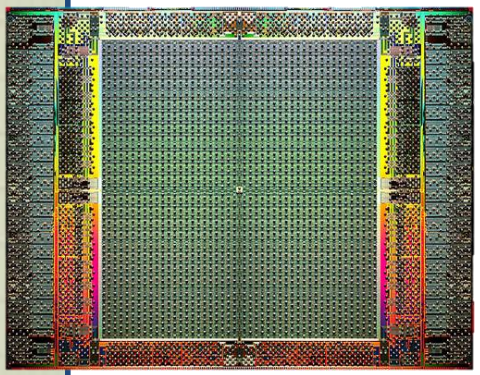
Custom Logic



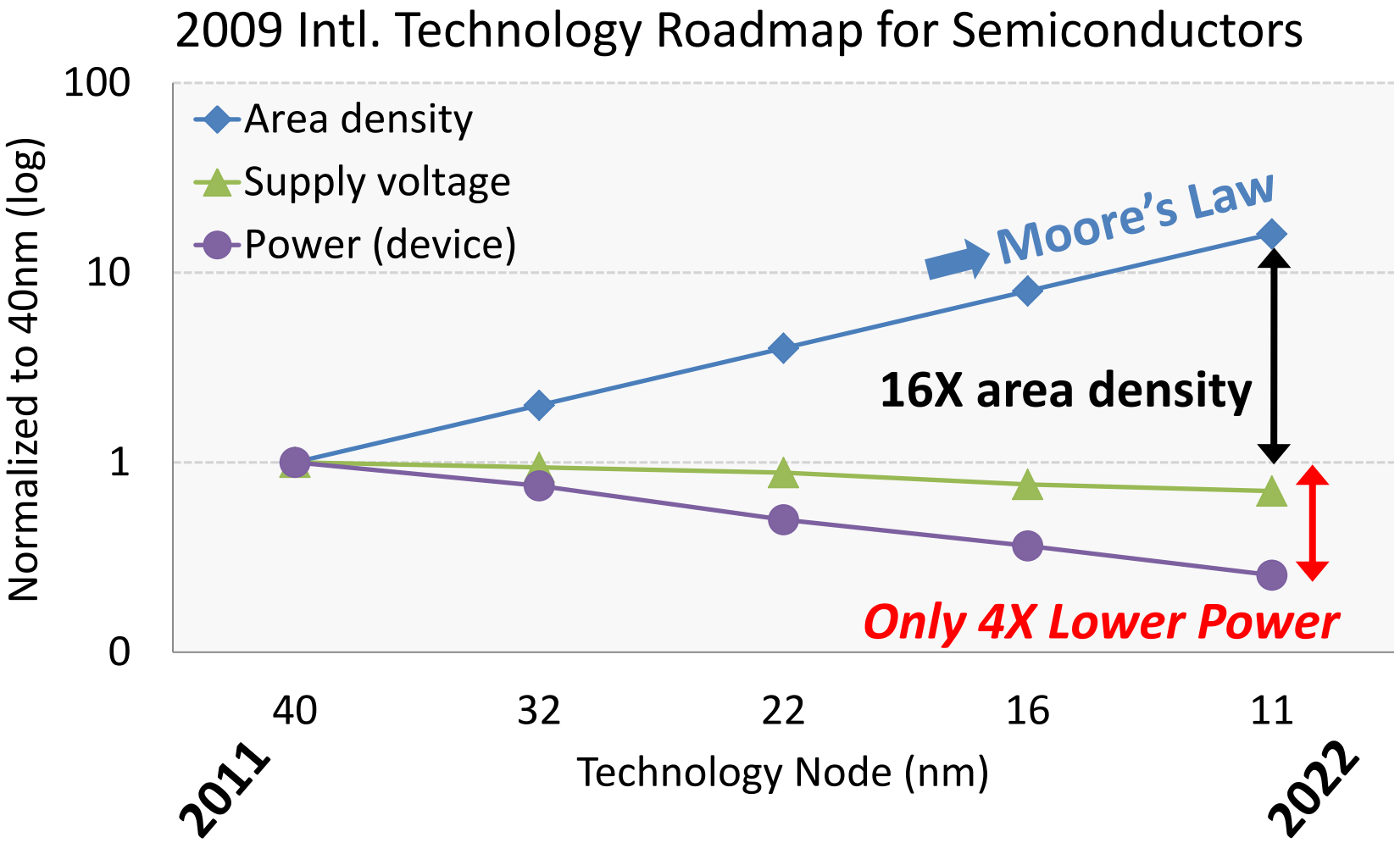


GPGPU

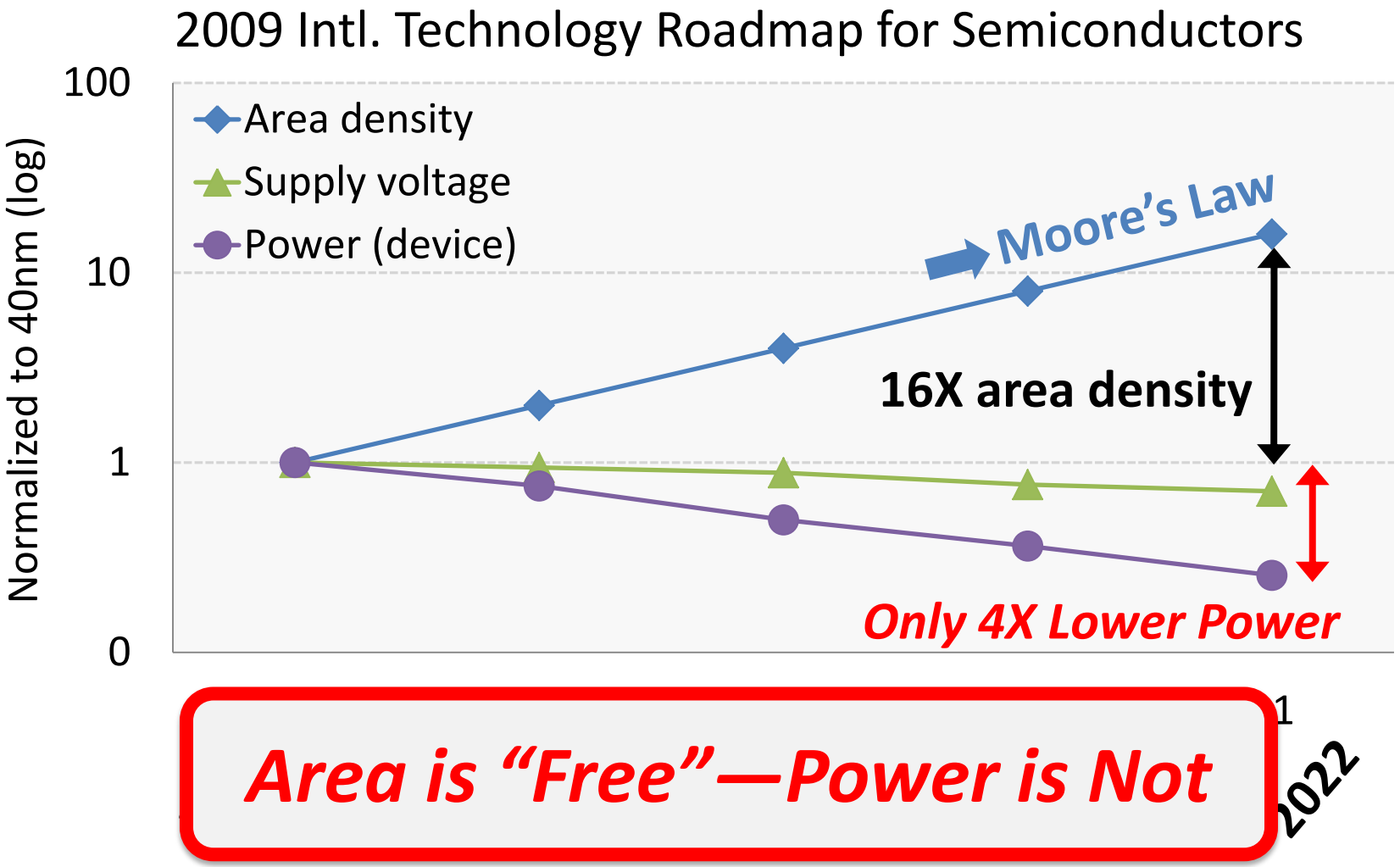
FPGA



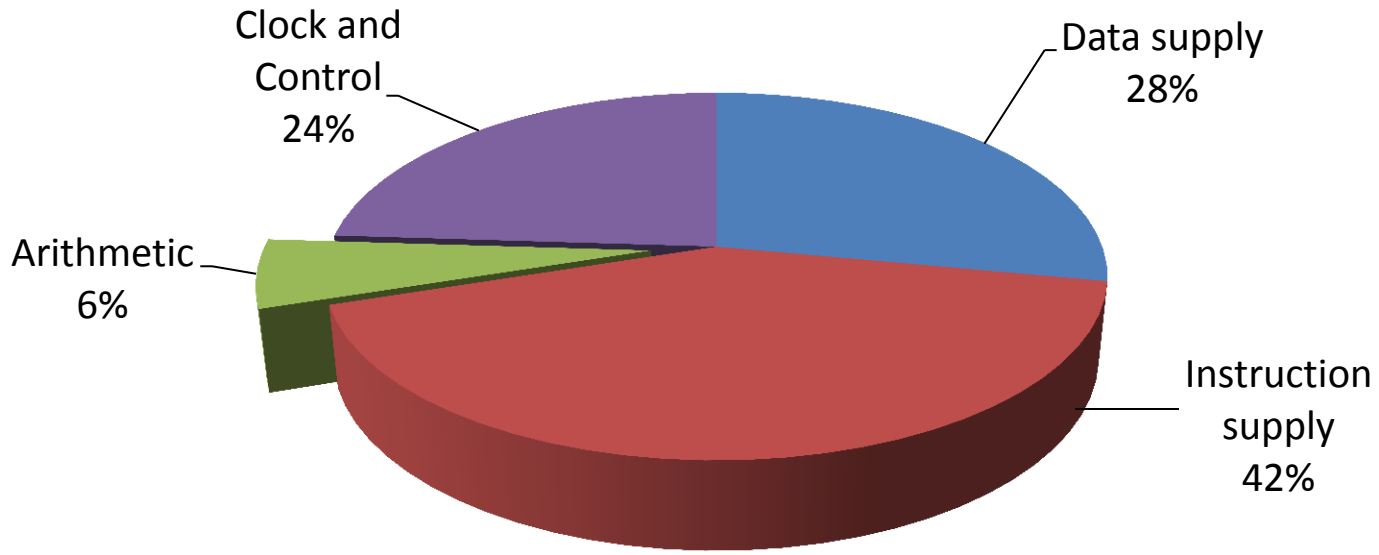
What We 'Know' About the Future



What We 'Know' About the Future



The Energy Efficiency of General Purpose Processors



Efficient Embedded Computing [Dally et al. 08]

Over 90% energy of general-purpose von Neumann processor is "overhead"

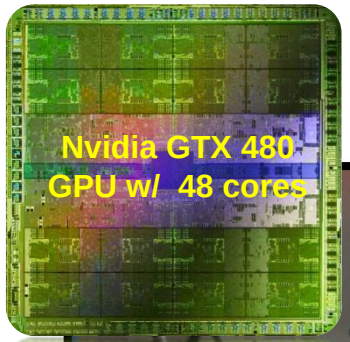
Overview

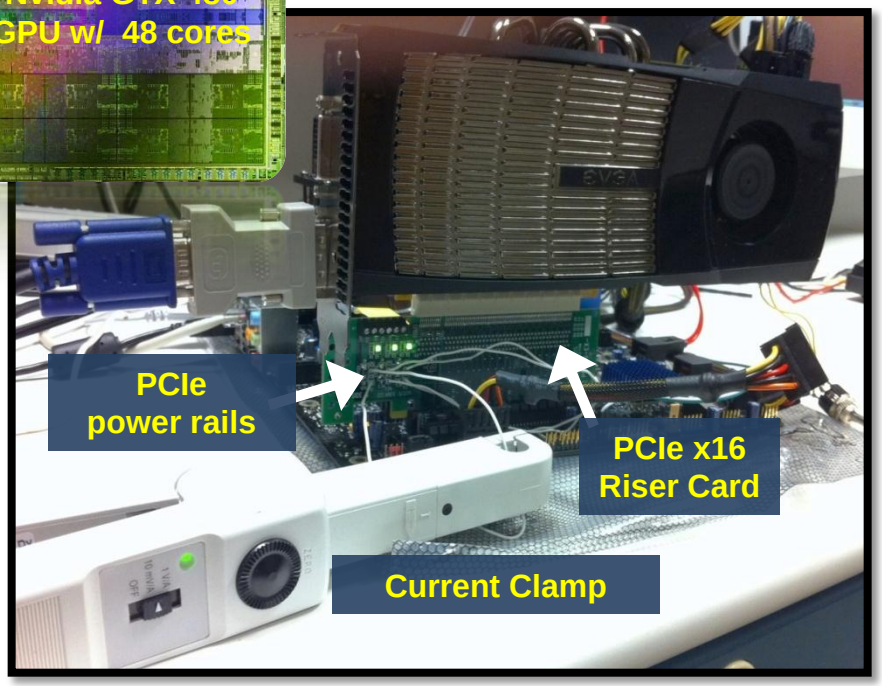
- Should the future look beyond general purpose?
- What is state of today's GPUs, FPGAs, and ASICs?

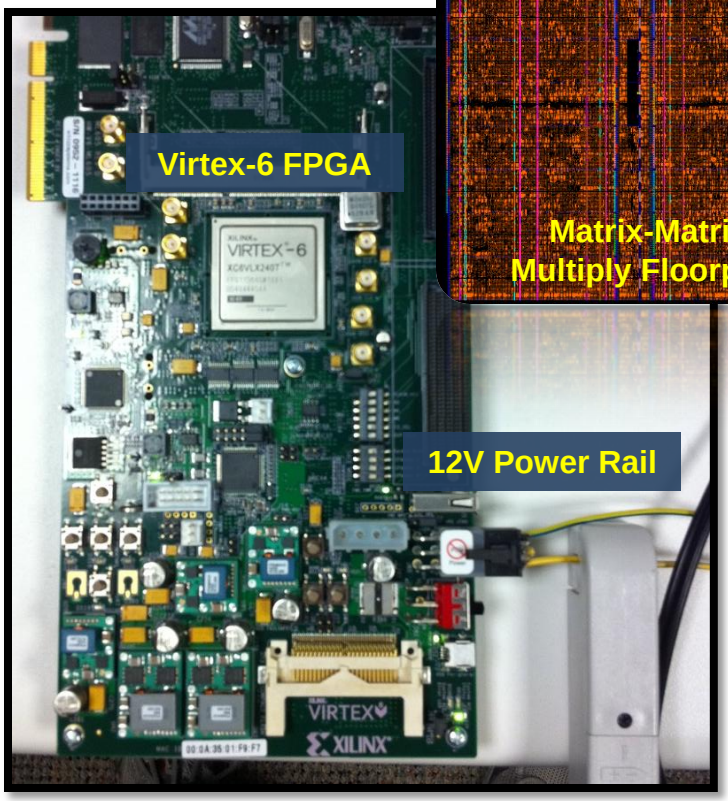
"GPUs are power hogs"	Answer: No
"FPGAs suck at floating point"	Answer: Not entirely
"ASICs are the best, right?"	Answer: …

- This talk
  - a perf and power study of today's computing alternatives
  - a model for future heterogeneous multicores

The Power and Performance of Today's GPUs, FPGAs, ASICs







GPUs, FPGAs, and All That

	CPU	GPUs			FPGA	ASIC
	Intel Core i7-960	Nvidia GTX285	Nvidia GTX480	ATI R5870	Xilinx V6-LX760	65nm Std. Cell
Year	2009	2008	2010	2009	2009	2007
Node	45nm	55nm	40nm	40nm	40nm	65nm
Die area	263mm^2	470mm^2	529mm^2	334mm^2	-	-
Clock rate	3.2GHz	1.5GHz	1.4GHz	1.5GHz	0.3GHz	-

Kernels	Characteristics
Matrix-Matrix Multiplication	Compute-intensive, simple memory access pattern
Fast Fourier Transform	Complex dataflow, low arithmetic intensity
Black-Scholes	Complex math operators, high arithmetic intensity

To Make It A Fair Game

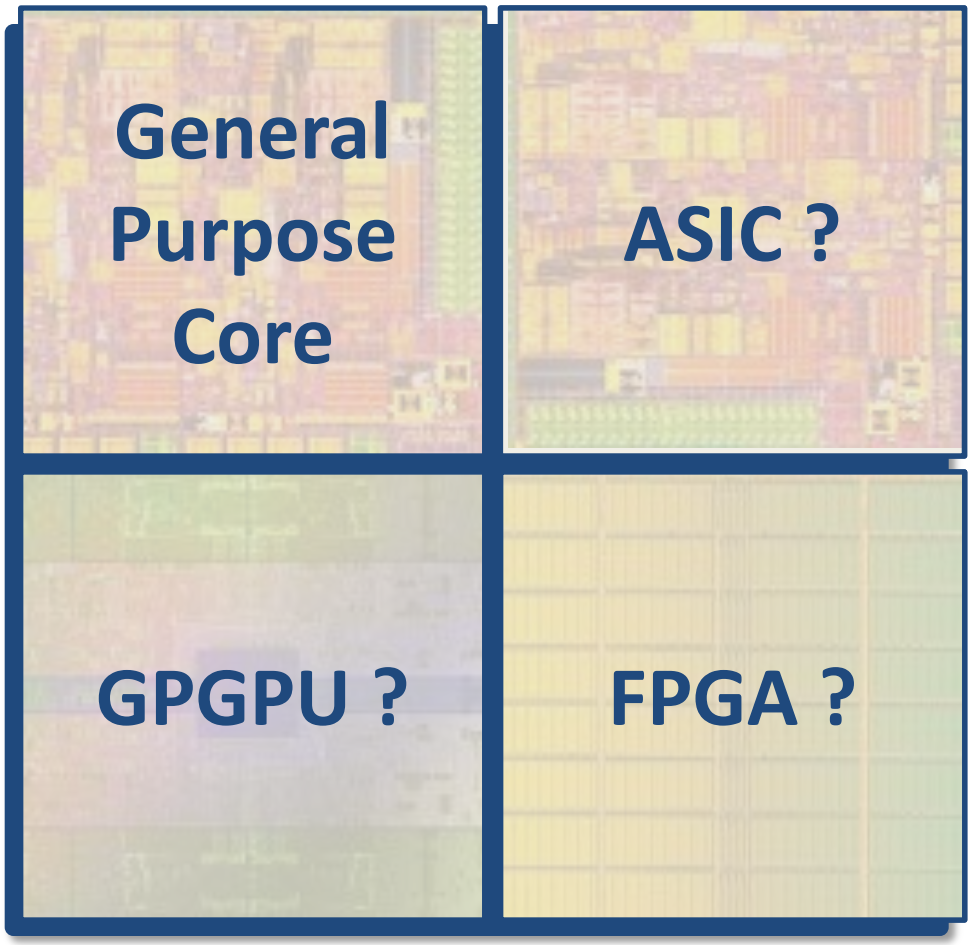
	Core i7	GTX285	GTX480	R5870	FPGA + ASIC
MMMult	MKL 10.2.3 multithreaded	CUBLAS 2.3	CUBLAS 3.1	CAL++	Hand-coded
FFT	Spiral.net multithreaded	CUFFT 2.3 3.0/3.1	CUFFT 3.0	-	Spiral.net
BScholes	PARSEC multithreaded	CUDA 2.3	-	-	Hand-coded

- Best-effort performance on all respective devices
- Single-precision floating point used for all kernels
- Power measurements isolated core from system
- All results normalized to same technology node

In-Core Performance and Energy

	Device	GFLOP/s actual	(GFLOP/s)/mm² norm. to 40nm	GFLOP/J norm. to 40nm
MMM	CPU-Core i7	96	0.50	**1.14**
	GPU-GTX480	541	1.28	**3.52**
	GPU-GTX285	425	2.40	**6.78**
	GPU-R5870	1491	5.95	**9.87**
	FPGA-LX760	204	0.53	**3.62**
	Same RTL in 65nm	694	19.28	**50.73**
		GFLOP/s	(GFLOP/s)/mm²	GFLOP/J
FFT-1024	CPU-Core i7	67	0.35	**0.71**
	GPU-GTX285	250	1.41	**4.2**
	GPU-GTX480	453	1.08	**4.3**
	GPU-R5870	-	-	**-**
	FPGA-LX760	380	0.99	**6.5**
	Same RTL in 65nm	952	239	**90**

Onward with a Future of Heterogeneous Multicores



A Model for Single-Chip Heterogeneous Multicores

- Baseline model: Amdahl's Law for Multicore [Hill & Marty 08]
- Assume designer has 'n' area resources to spend

Base Core Equivalent

Consumes 1 unit of area

$Perf_{BCE} = 1$

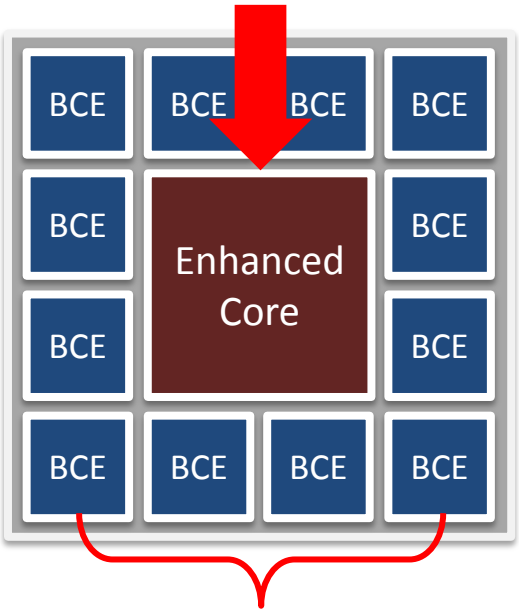
Enhanced Core

Consumes *r* units of area

$Perf_{enh}(r) \approx \sqrt{r}$ *(Pollack's Law)*

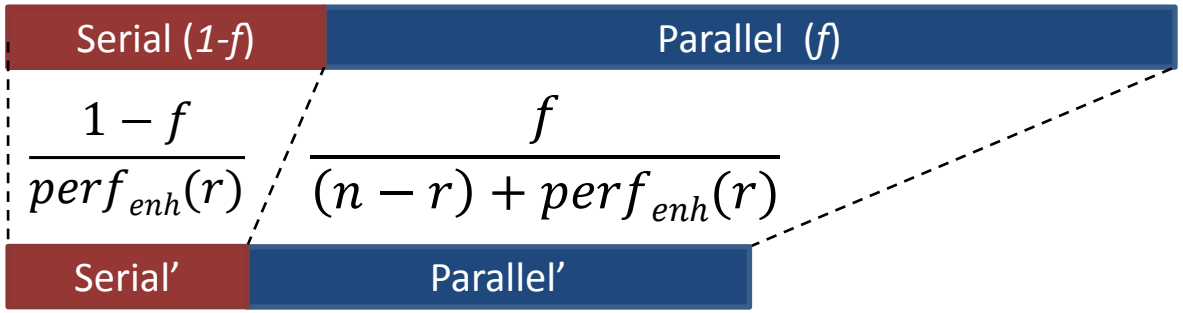
Asymmetric Multicore

r-sized enhanced core



(n – r) BCE cores

Original execution time on 1 BCE

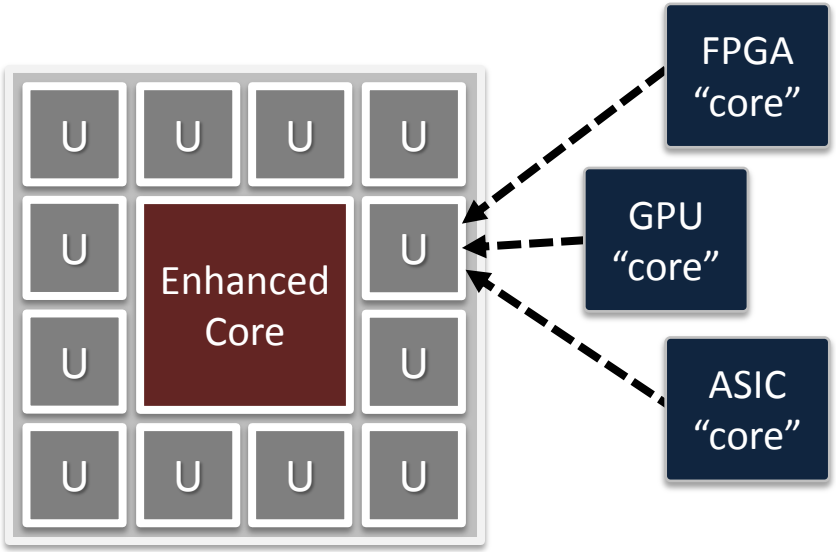


Assumptions:
(1) fraction 'f' perfectly parallelizable, '1-f' totally serial
(2) no scheduling/communication/sync/etc overheads

$$Speedup = \frac{1}{\frac{1-f}{perf_{enh}(r)} + \frac{f}{(n-r)+perf_{enh}(r)}}$$

Extending Amdahl's for Heterogeneous Multicore

- Define U-core as BCE-equivalent area of FPGA/GPU/ASIC
- U-cores can only execute parallel kernels (MMM, FFT, etc)



Each U-core characterized by:

- *performance μ* ***and***
- *power ϕ relative to BCE core*

(BCE gives 1 unit performance & consumes 1 unit power)

Kernel ϕ and μ Values

		MMM	Black-Scholes	FFT-1024
Nvidia GTX285	Φ	0.7	0.6	0.6
	μ	3.4	17	2.9
Nvidia GTX480	Φ	0.8	-	0.5
	μ	1.8	-	2.2
ATI R5870	Φ	1.3	-	-
	μ	8.5	-	-
Xilinx LX760	Φ	0.3	0.3	0.3
	μ	0.8	6	2
Custom Logic	Φ	0.8	5	5
	μ	27	480	490

Nominal BCE based on an Intel Atom in-order processor, 26mm^2 in a 45nm process

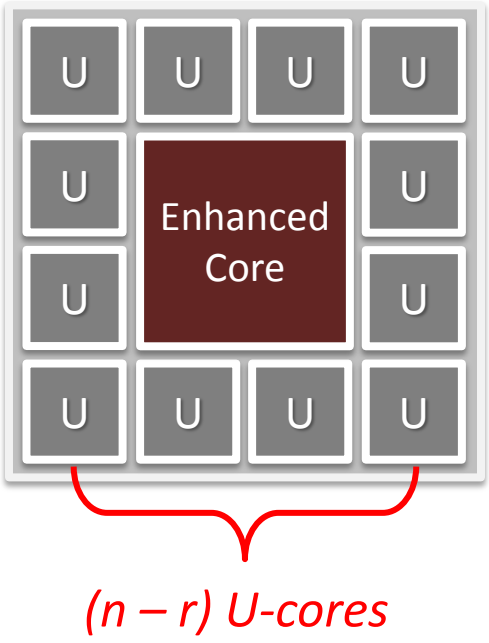
Kernel ϕ and μ Values

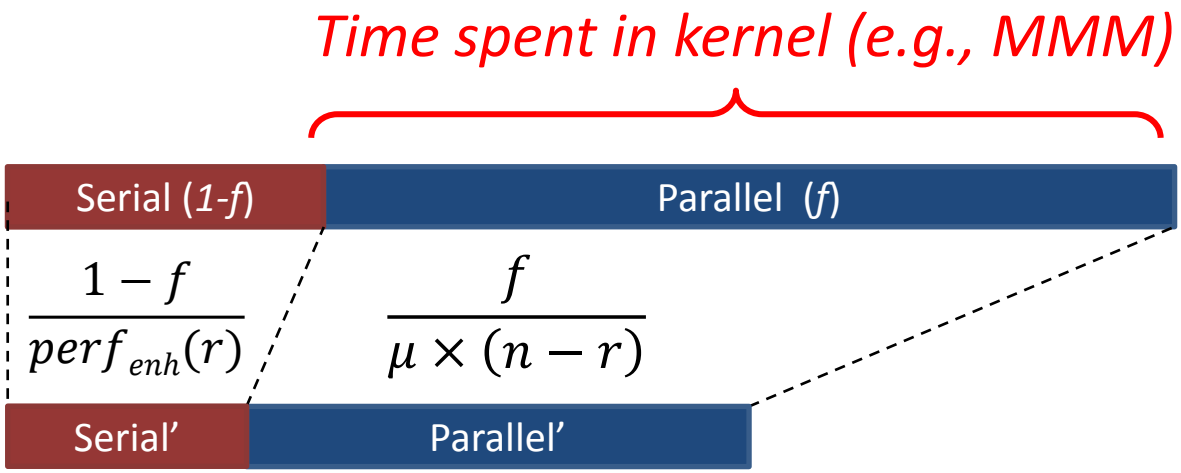
		MMM	Black-Scholes	FFT-1024
Nvidia GTX285	Φ	0.7	0.6	0.6
	μ	3.4	17	2.9
Nvidia GTX480	Φ	0.8		
	μ	1.8		
ATI R5870	Φ	1.3		
	μ	8.5		
Xilinx LX760	Φ	0.3		
	μ	0.8	6	2
Custom Logic	Φ	0.8	5	5
	μ	27	480	490

On equal area basis, 3.4 performance at 0.7X power relative to a BCE

Nominal BCE based on an Intel Atom in-order processor, 26mm^2 in a 45nm process

Extending Amdahl's for Heterogeneous Multicore



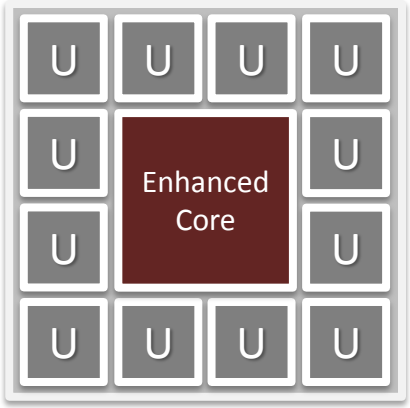


$$Speedup = \frac{1}{\frac{1-f}{perf_{enh}(r)} + \frac{f}{\mu \times (n-r)}}$$

μ = relative U-core performance to BCE core

Modeling Power and Bandwidth Budgets

$$Speedup = \frac{1}{\frac{1-f}{perf_{enh}(r)} + \frac{f}{\mu \times (n-r)}}$$

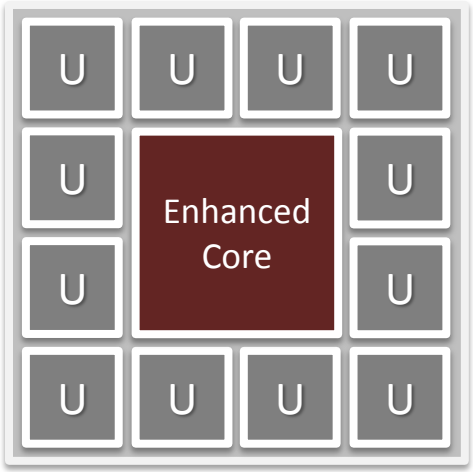


- The above is based on area alone
- Power or bandwidth budget limits the usable die area (*n*)
  - if *P* is total power budget expressed as a multiple of a BCE's power, then usable U-core area constrained by $\phi \times (n-r) \leq P$
  - if *B* is total memory bandwidth expressed also as a multiple of BCEs, then usable U-core area constrained by $\mu \times (n-r) \leq B$

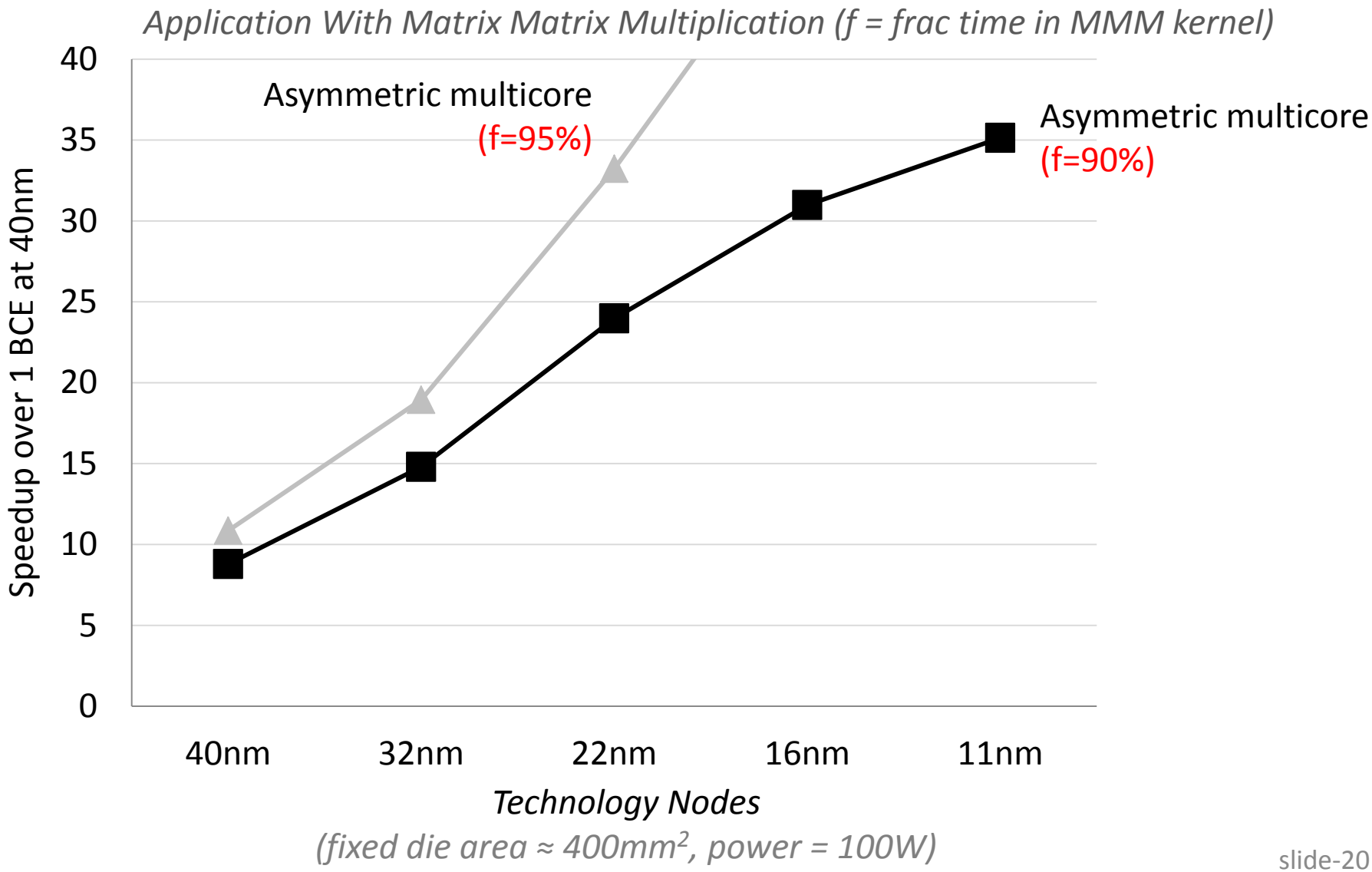
Combine Model with ITRS Trends

Year	2011	2013	2016	2019	2022
Technology	40nm	32nm	22nm	16nm	11nm
Core die budget (mm^2)	432	432	432	432	432
Normalized area (BCE)	19	37	75	149	298
Core power (W)	100	100	100	100	100
Bandwidth (GB/s)	180	198	234	234	252
Relative power per device	1X	0.75X	0.5X	0.36X	0.25X

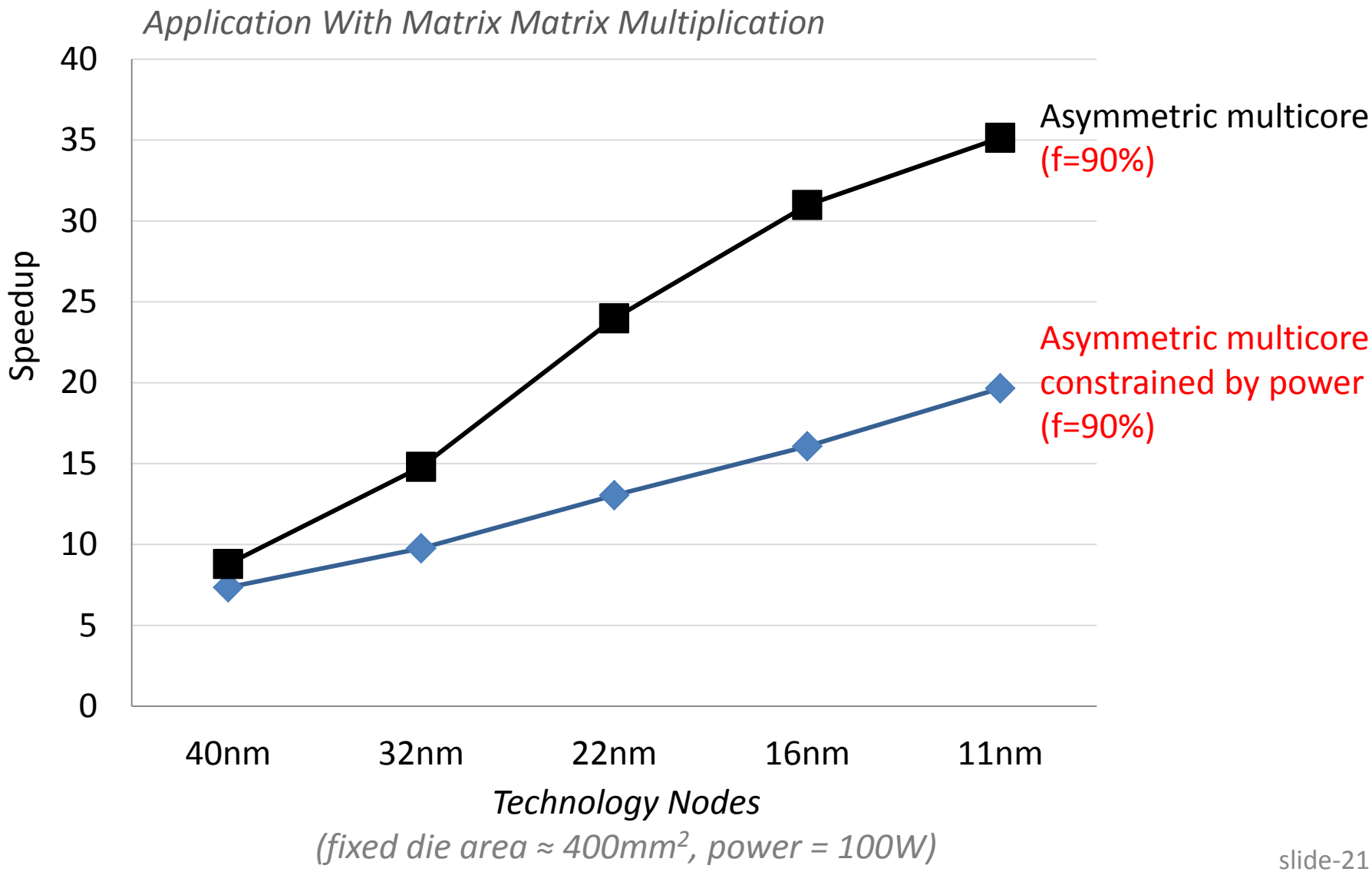
- 2011 parameters reflect high-end systems today; future parameters extrapolated from ITRS road map
- 432mm^2 populated by an optimally sized Fast Core and U-cores of choice



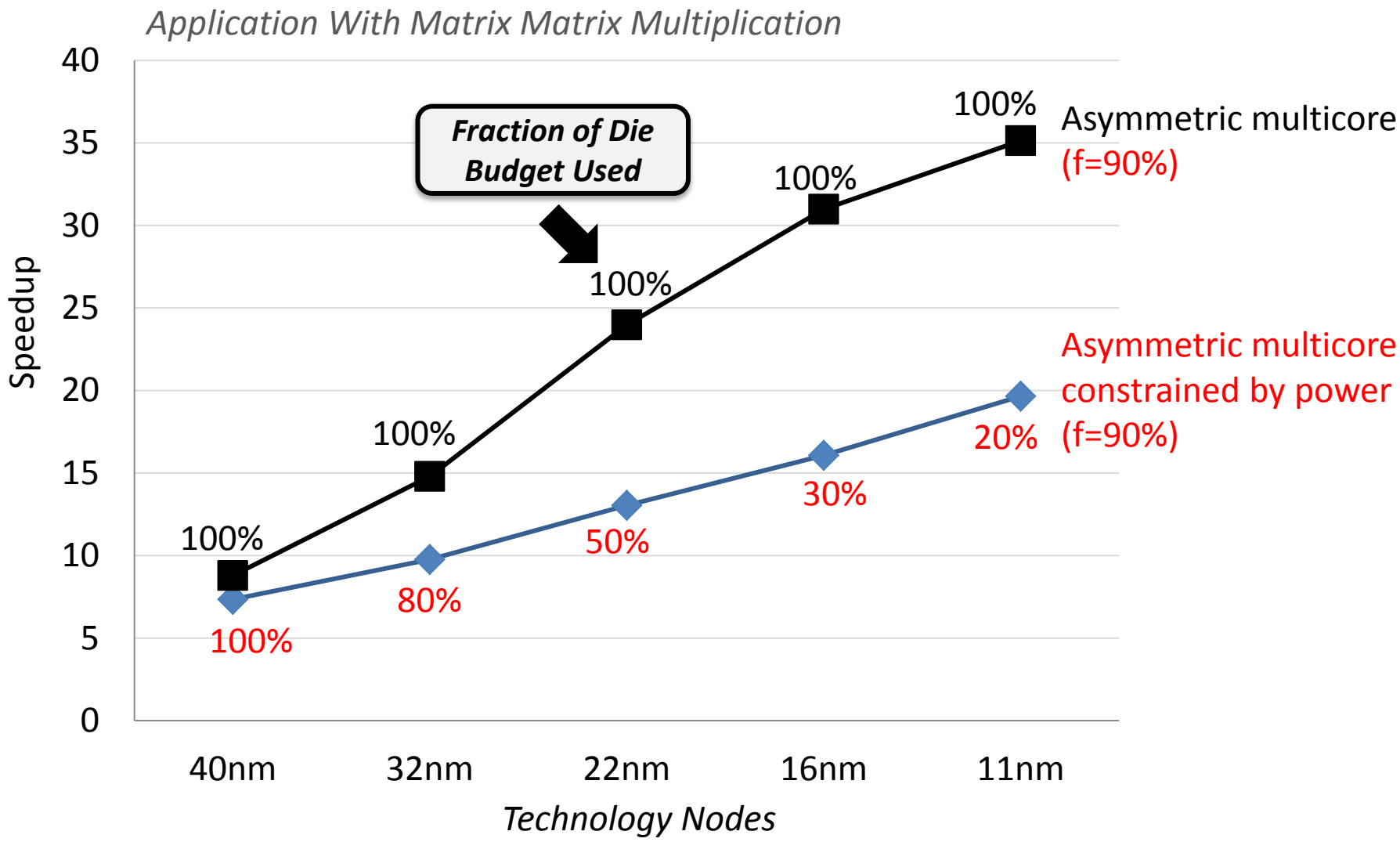
Asymmetric Multicore Trends



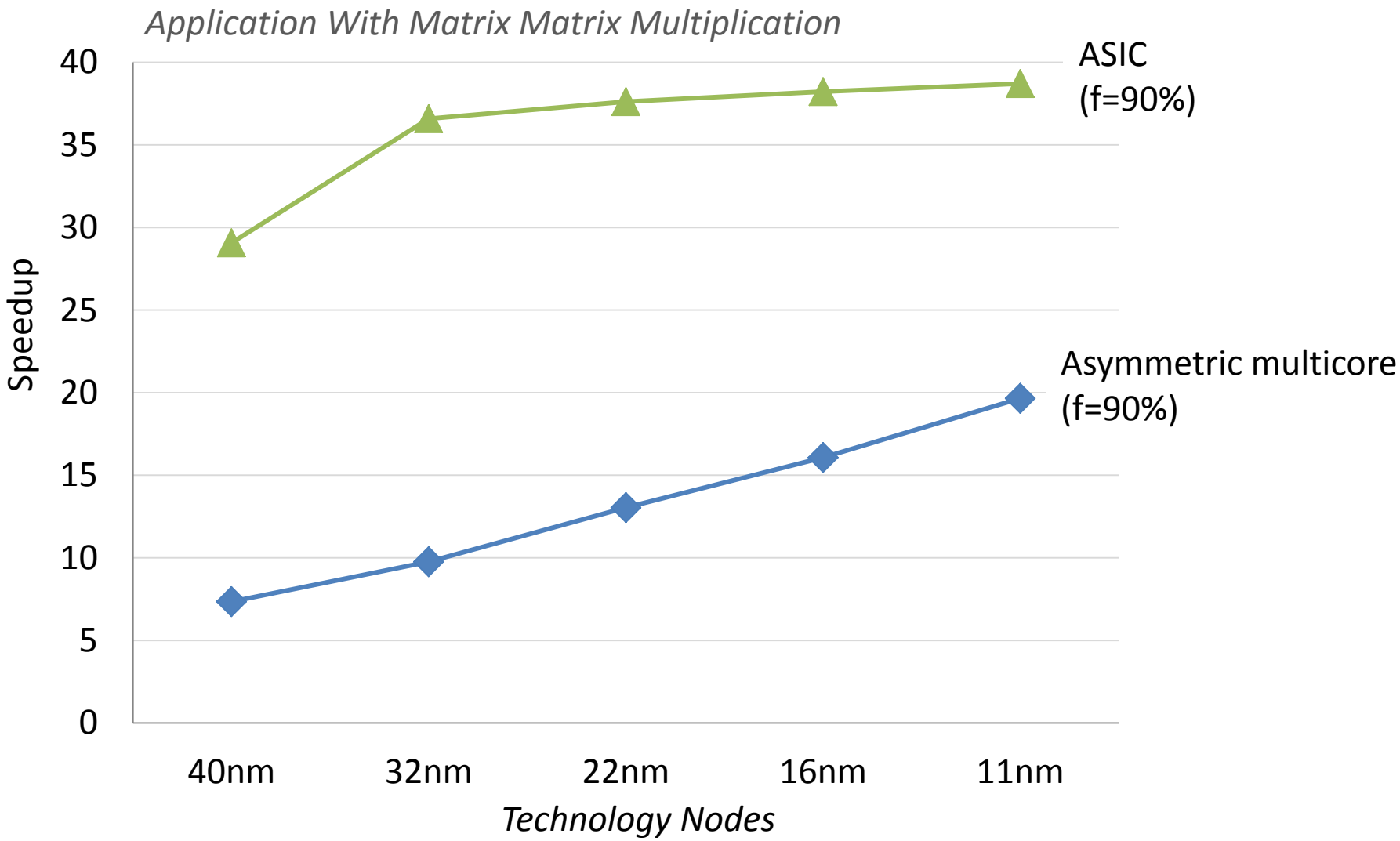
Asymmetric Multicore Trends



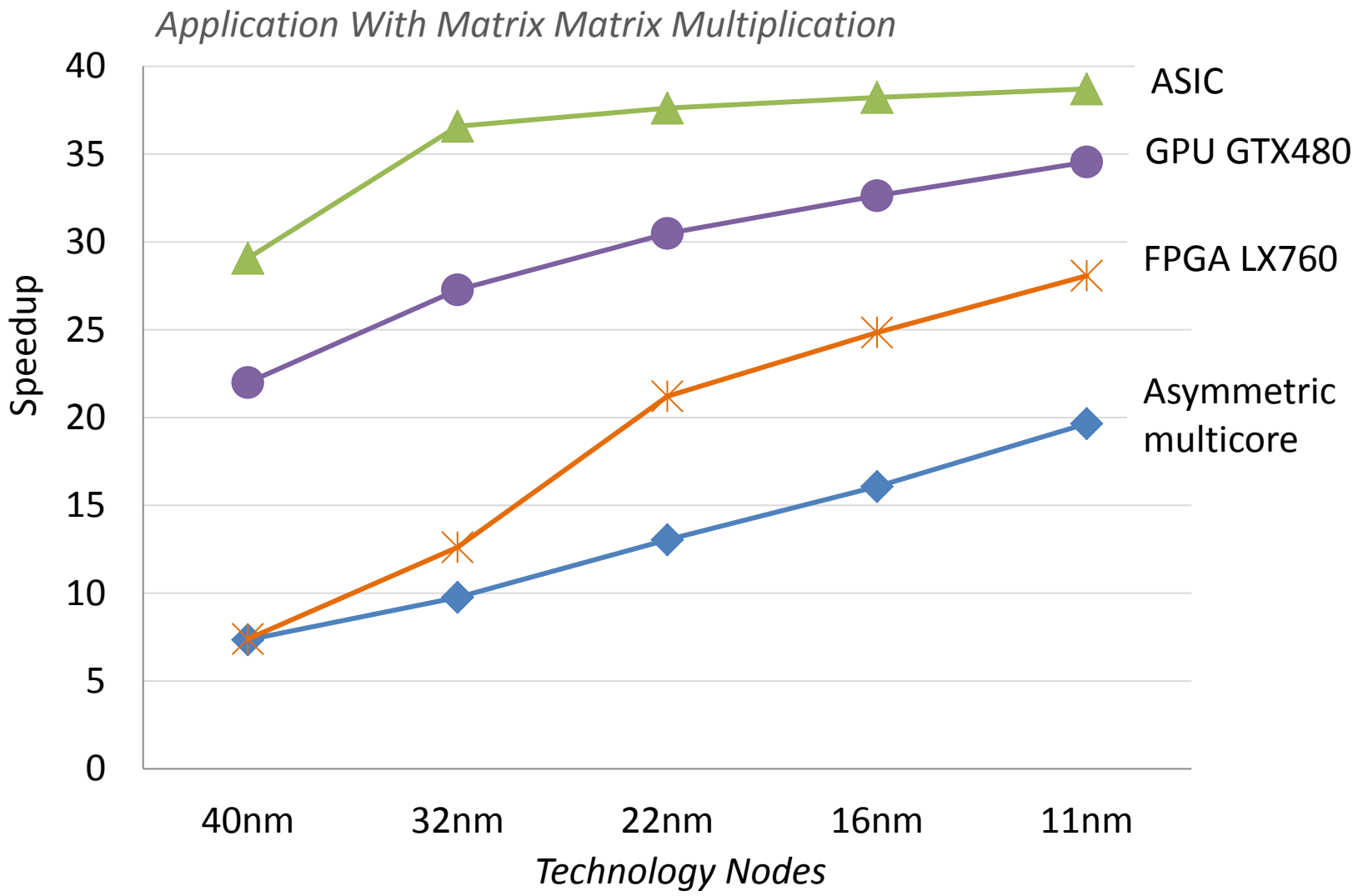
Asymmetric Multicore Trends



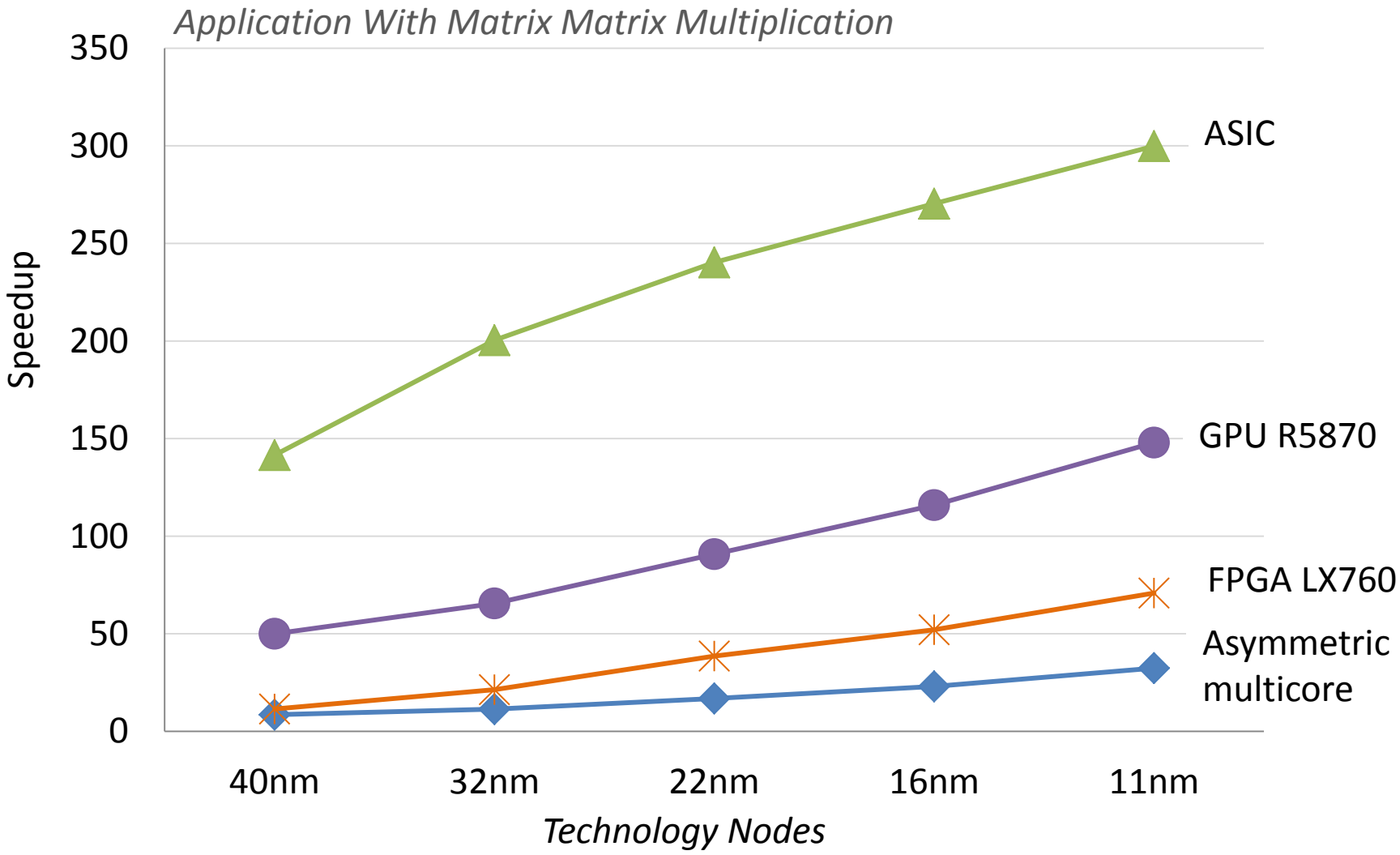
Asymmetric versus ASIC



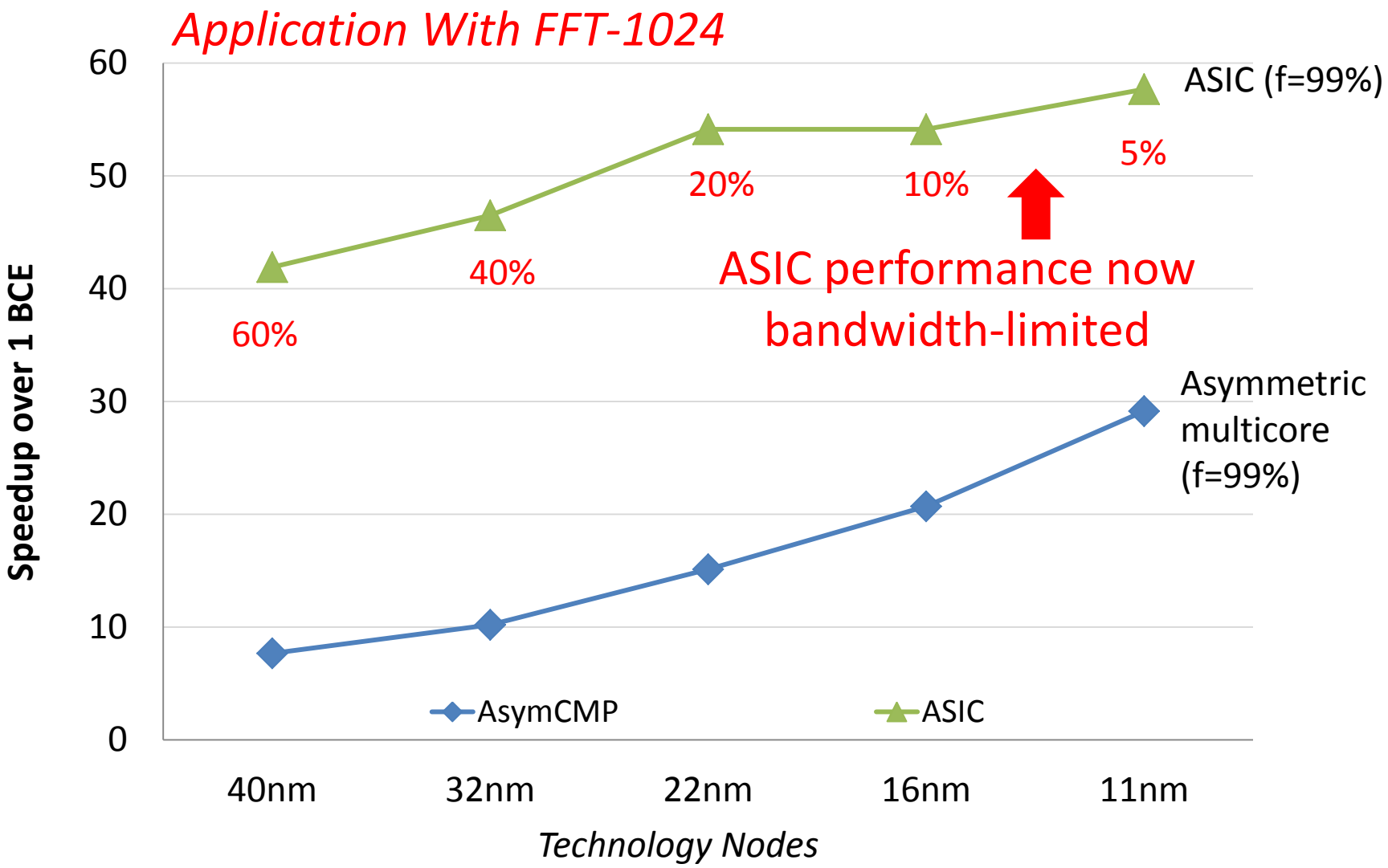
How Do GPU/FPGA Compare?



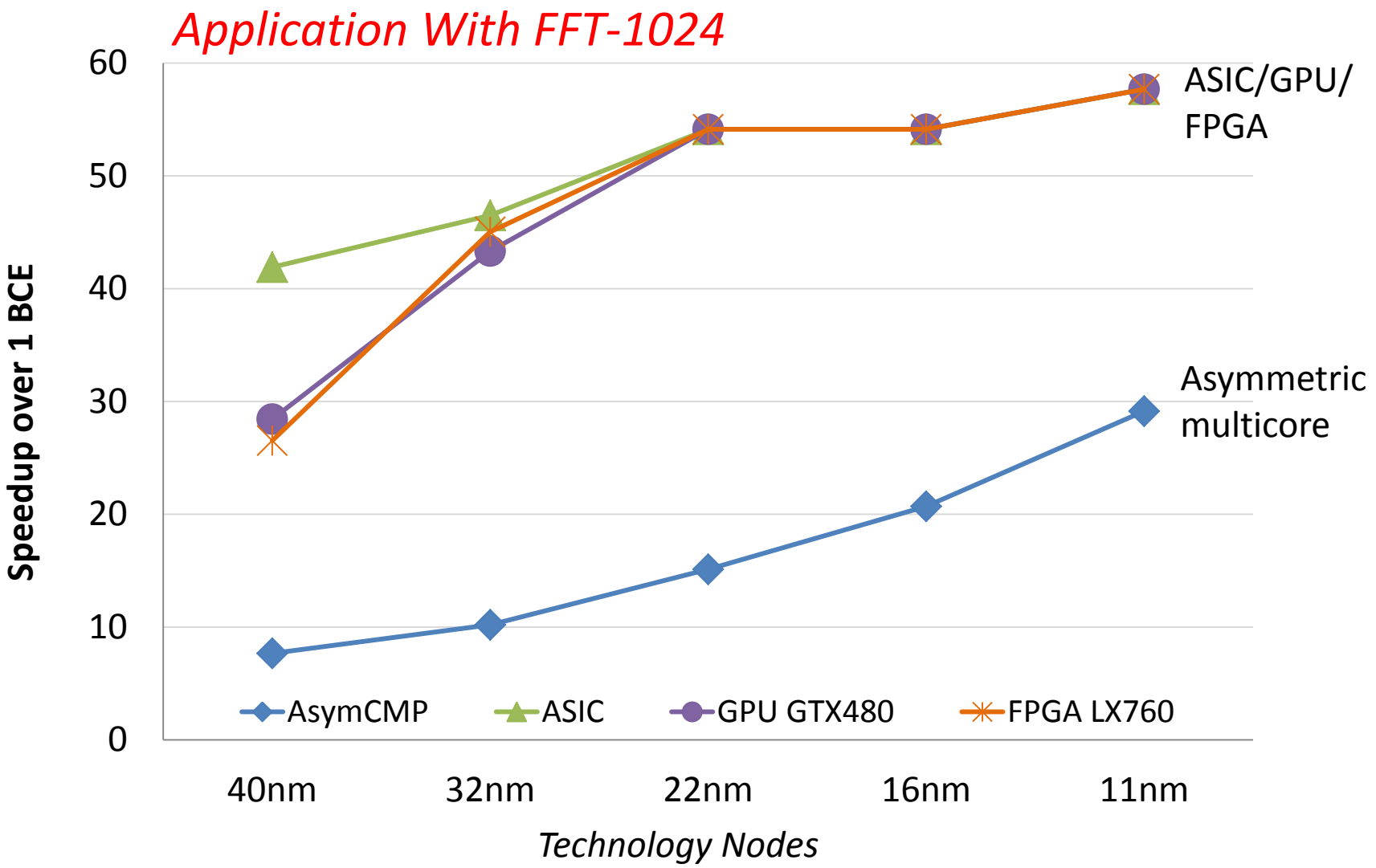
What if 'f' increased to 99%?



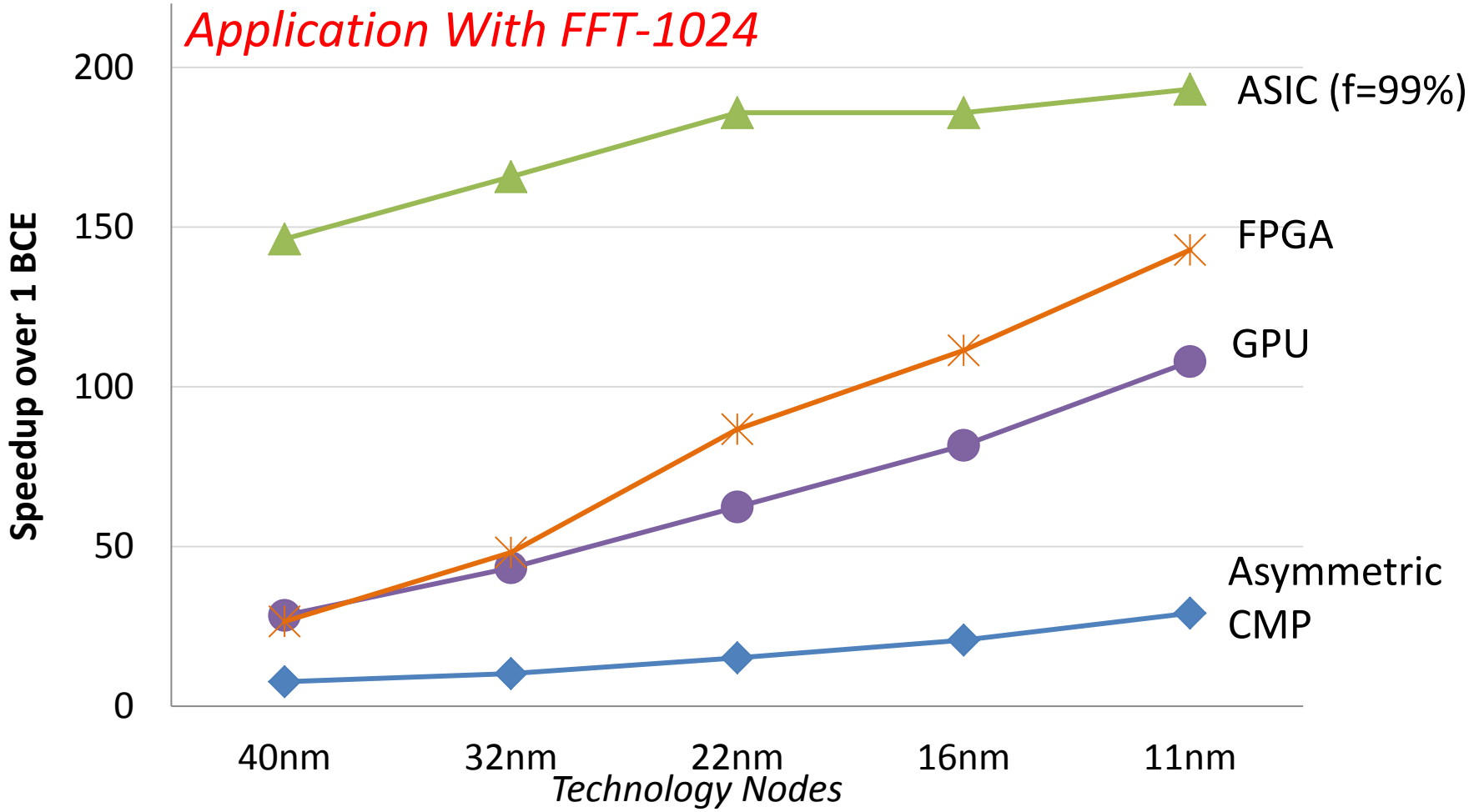
Impact of Mem Bandwidth?



Impact of Mem Bandwidth?



What If We Had 3D Mem Stacking? (BW>1TB/s)



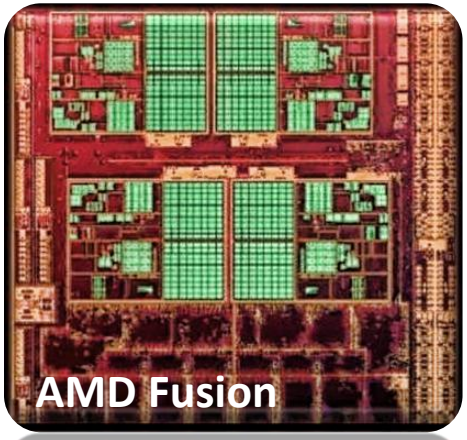
Conclusions

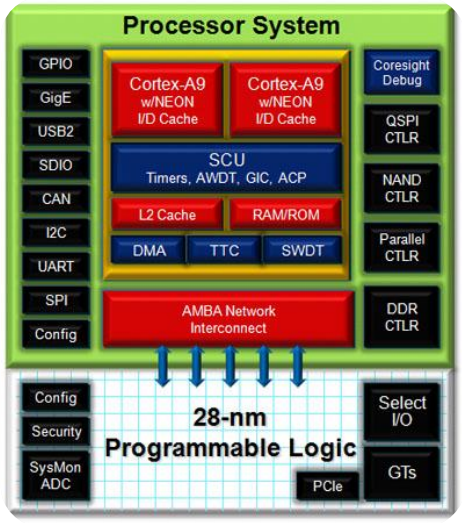
- **The 'best' choice of U-cores not always obvious**
  - technology trends have major impact on relative merits
  - must think about power, bandwidth, and parallelism together
- **U-cores help performance *only if***
  - significant fraction amenable to acceleration, *and*
  - adequate BW to sustain acceleration *3D-stacked memory could help!*
- **Programmable GPUs/FPGAs can keep up with ASICs**
  - ASICs need substantial parallelism and/or bandwidth to make sense
  - maximizing energy efficiency can have diminishing returns on perf

ASICs still best if energy reduction is objective (details in paper)

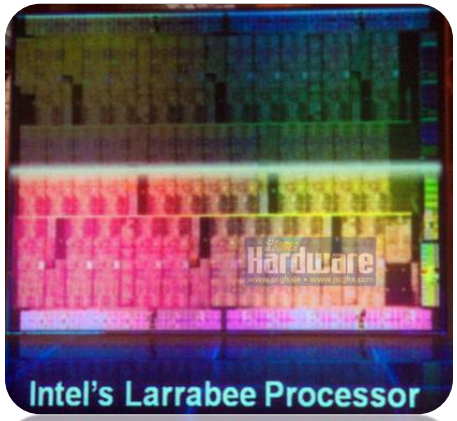
Should the Future Include Custom Logic, FPGAs, and GPGPUs?

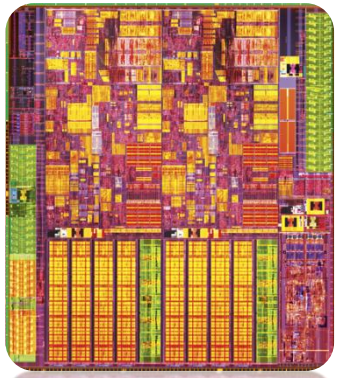
Still A Question?





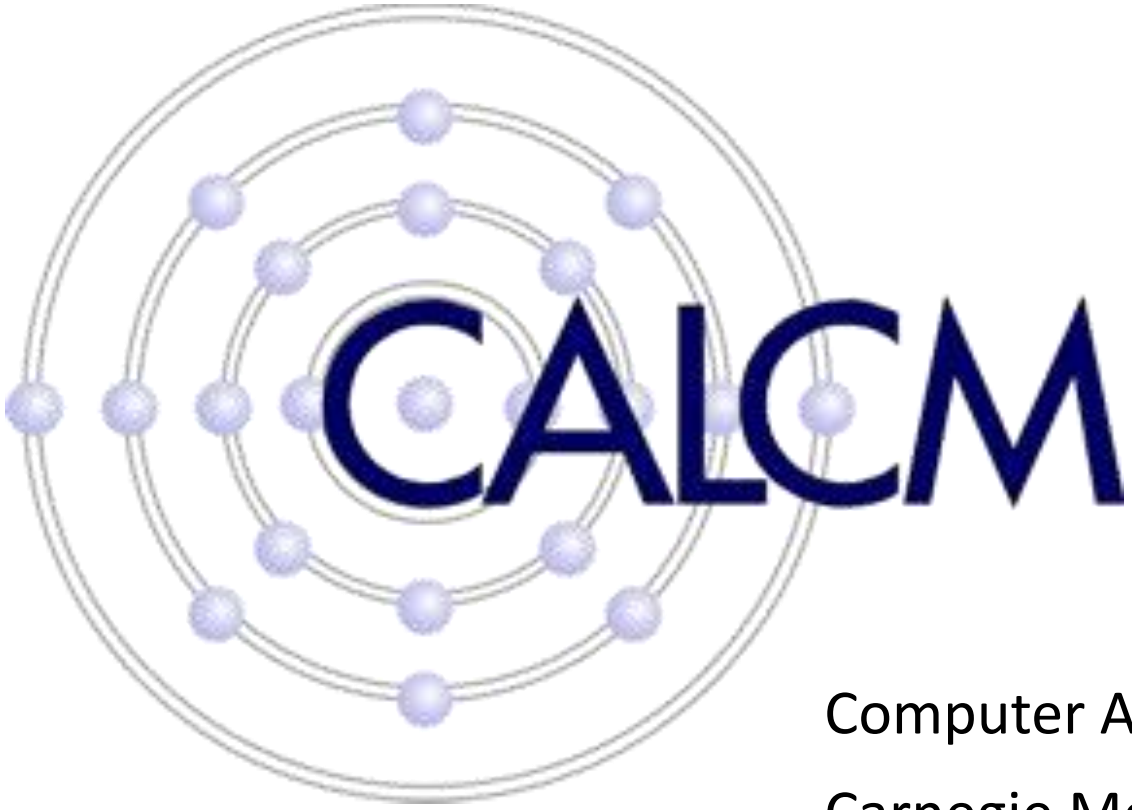
Xilinx FPGA + ARM Multicore





Intel Core i5
(CPU + GPU integrated)

Thank You!



Computer Architecture Lab (CALCM)

Carnegie Mellon University

http://www.ece.cmu.edu/CALCM