On the Communication Complexity of 3D FFTs and its Implications for Exascale

Kent Czechowski, Chris McClanahan, Casey Battaglino, Kartik Iyer,
P.-K. Yeung, and Richard Vuduc

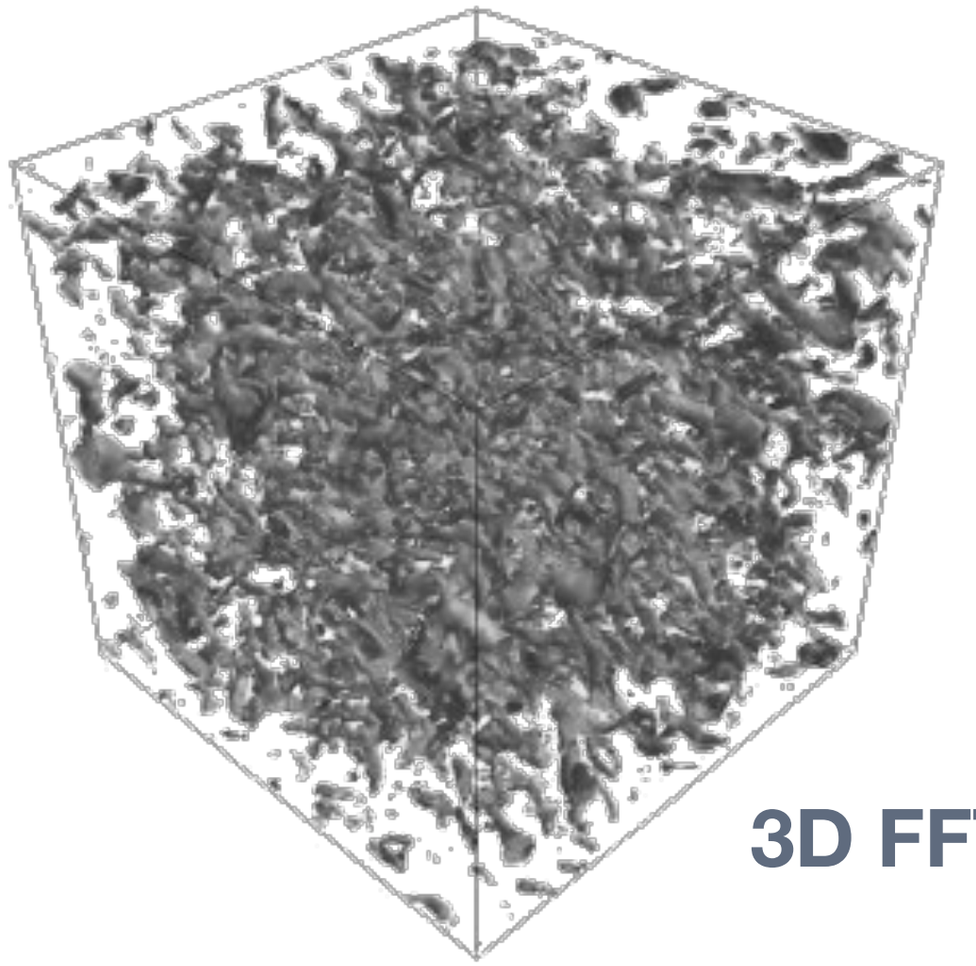
Exascale-ability

Today

$N=4096^3$

12.3×10^{12} Flops

1.1 TB of Data



3D FFT

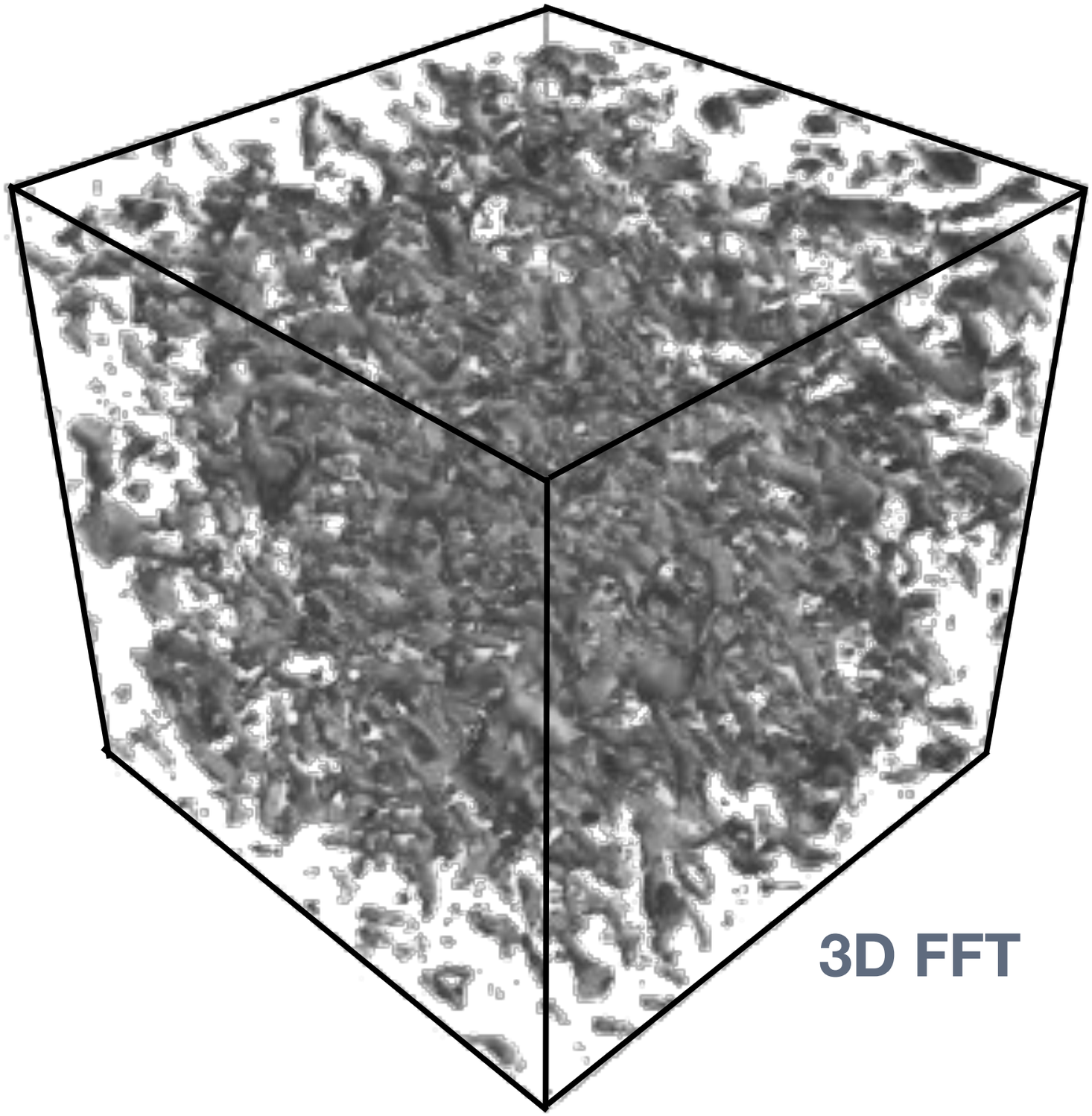
Exascale-ability

Tomorrow

$N=131,072^3$

$.574\times10^{18}$ Flops

36 PB of Data

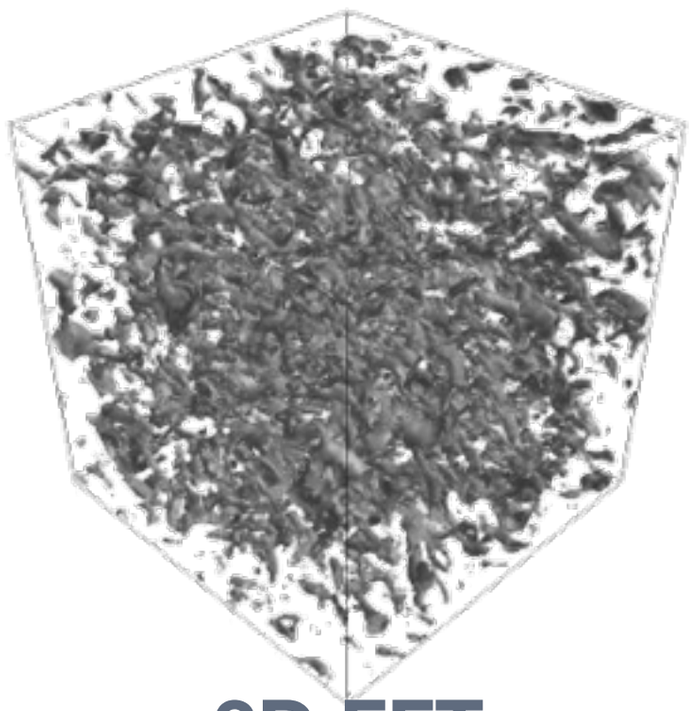


vs.

Swim lane 1

Swim lane 2

\+

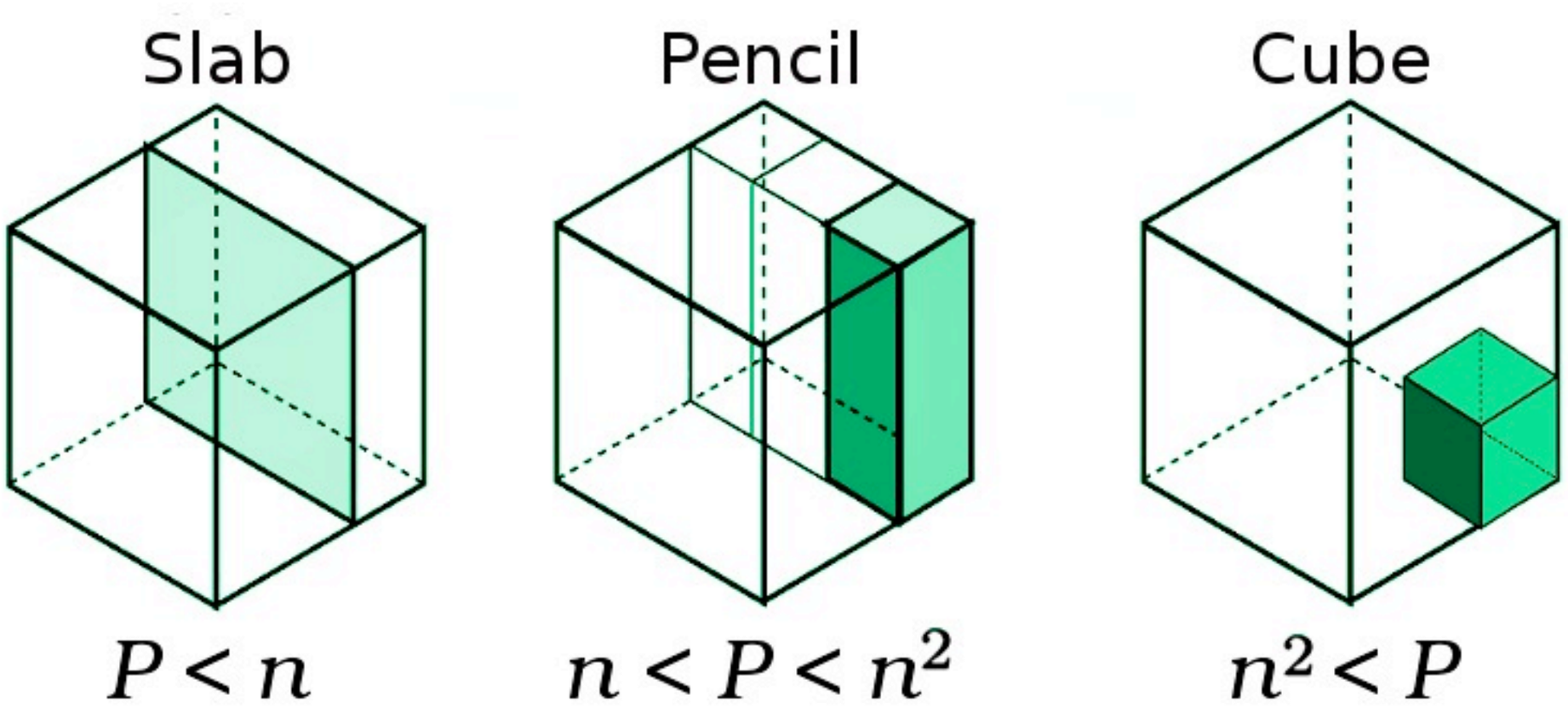


3D FFT

= ?

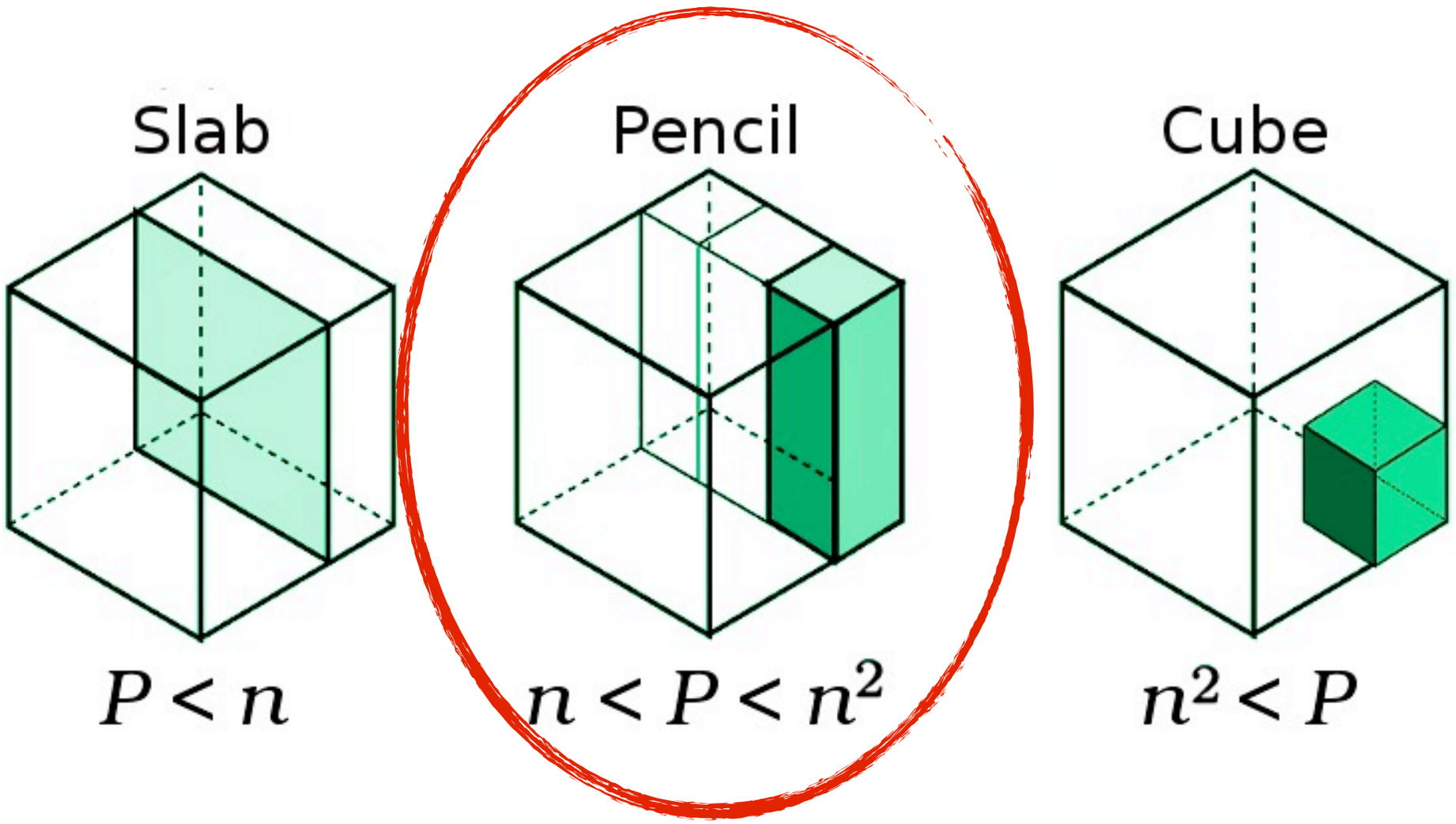
Performance Model

3D FFT Decompositions



Problem size N=n×n×n

Pencil Decomposition



Problem size N=n×n×n

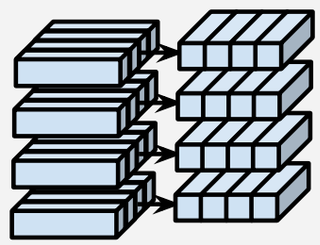
Distributed 3D FFT - Performance Model

3D FFT Using the Pencil Decomposition of the Transpose Method

Computation Phase #1

1D FFT in the X-direction

Communication Phase #1

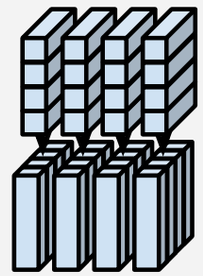


X → Y Transpose

Computation Phase #2

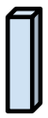
1D FFT in the Y-direction

Communication Phase #2



Y → Z Transpose

Computation Phase #3



1D FFT in the Z-direction

Distributed 3D FFT - Performance Model

Computation Phase

1D FFT in the X-direction

Each node computes n^2/p 1D FFTs of size n

Arithmetic Computation Time

$$T_{\text{flops}} = 3 \times \frac{n^2}{P} \times \frac{5n \log n}{C_{\text{node}}}$$

Nodes: P
Compute Throughput: C_{node}

Memory Access Time

$$T_{\text{mem}} \approx 3 \times \frac{n^2}{P} \cdot \frac{A \times n(\max(\log_Z n, 1.0))}{\beta_{\text{mem}}}$$

Cache Capacity: Z
Memory BW: β_{mem}

Distributed 3D FFT - Performance Model

Computation Phase

1D FFT in the X-direction

Each node computes n^2/p 1D FFTs of size n

Arithmetic Computation Time

$$T_{\text{flops}} = 3 \times \frac{n^2}{P} \times \frac{5n \log n}{C_{\text{node}}}$$

Memory Access Time

$$T_{\text{mem}} \approx 3 \times \frac{n^2}{P} \cdot \frac{A \times n(\max(\log_Z n, 1.0))}{\beta_{\text{mem}}}$$

Nodes: P
Compute Throughput: C_{node}

Cache Capacity: Z
Memory BW: β_{mem}

Distributed 3D FFT - Performance Model

Computation Phase

1D FFT in the X-direction

Each node computes n^2/p 1D FFTs of size n

Arithmetic Computation Time

$$T_{\text{flops}} = 3 \times \frac{n^2}{P} \times \frac{5n \log n}{C_{\text{node}}}$$

Memory Access Time

$$T_{\text{mem}} \approx 3 \times \frac{n^2}{P} \cdot \frac{A \times n(\max(\log_Z n, 1.0))}{\beta_{\text{mem}}}$$

$\Theta\left(1 + (n/L)(1 + \log_Z n)\right)$

Lower bound (Frigo 1999)

Nodes: P

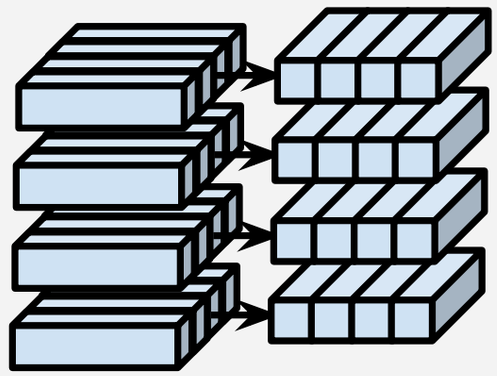
Compute Throughput: C_{node}

Cache Capacity: Z

Memory BW: β_{mem}

Distributed 3D FFT - Performance Model

Communication Phase



X → Y Transpose

√p-node All-to-All communications

Network Time

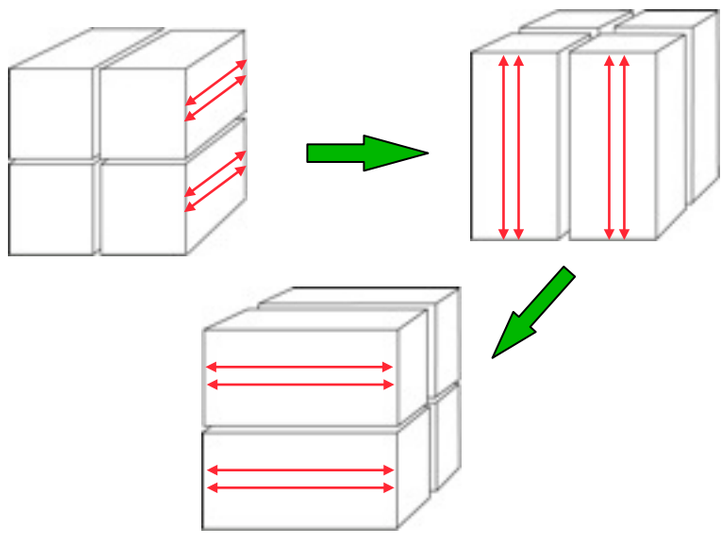
$$T_{\text{net}} \approx 2 \times \frac{n^3}{P^{\frac{2}{3}} \beta_{\text{link}}}$$

Nodes: P Network BW: β_{link}

Validation

3D FFT Software

Distributed 3D FFT Framework



p3dfft

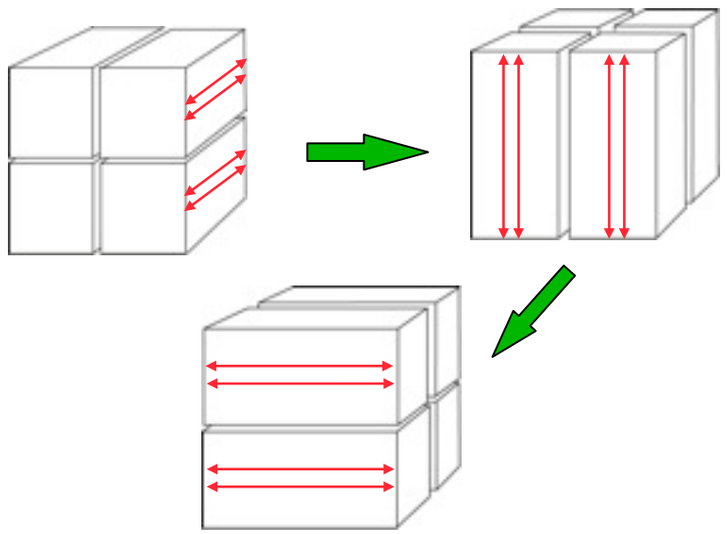
+

Optimized 1D FFT Library

FFTW
ESSL
MKL

3D FFT Software

Distributed 3D FFT Framework



p3dfft

+

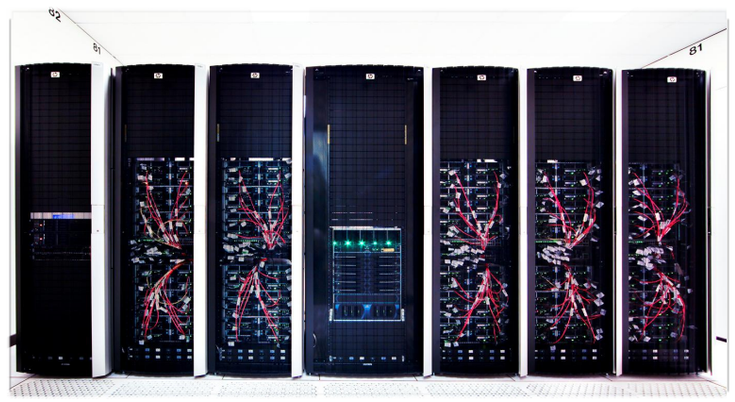
Optimized 1D FFT Library

FFTW
ESSL
MKL
CUFFT

Test Machines

Hopper
6,392 Nodes
Opteron 6100 CPU

Processor Peak: 50.4 GF/s
Cores: 6
Memory BW: 21.3 GB/s
Fast Memory: 6 MB
Link BW: 10 GB/s

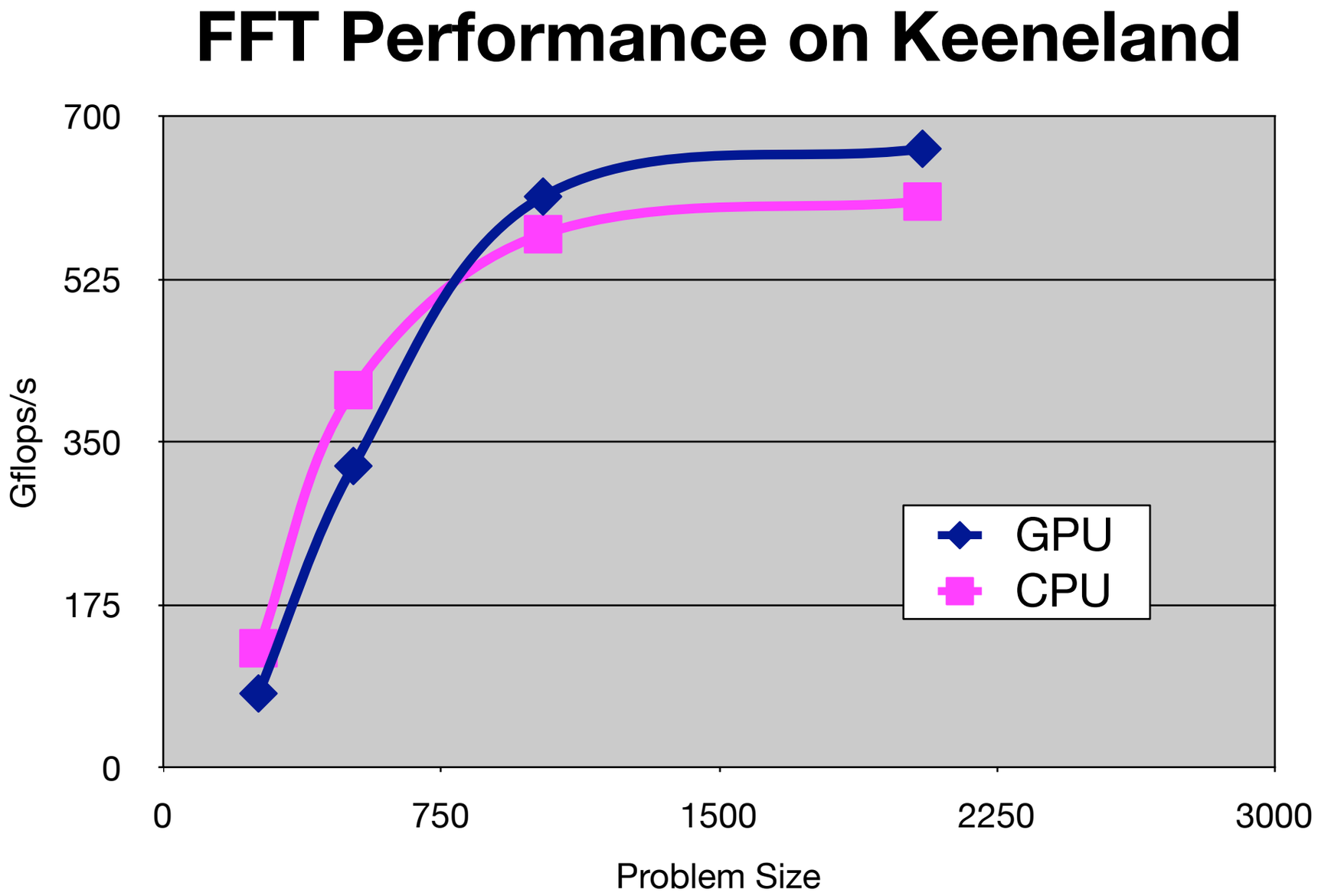


Keeneland
120 Nodes (3xGPUs per node)
Tesla M2070 GPU

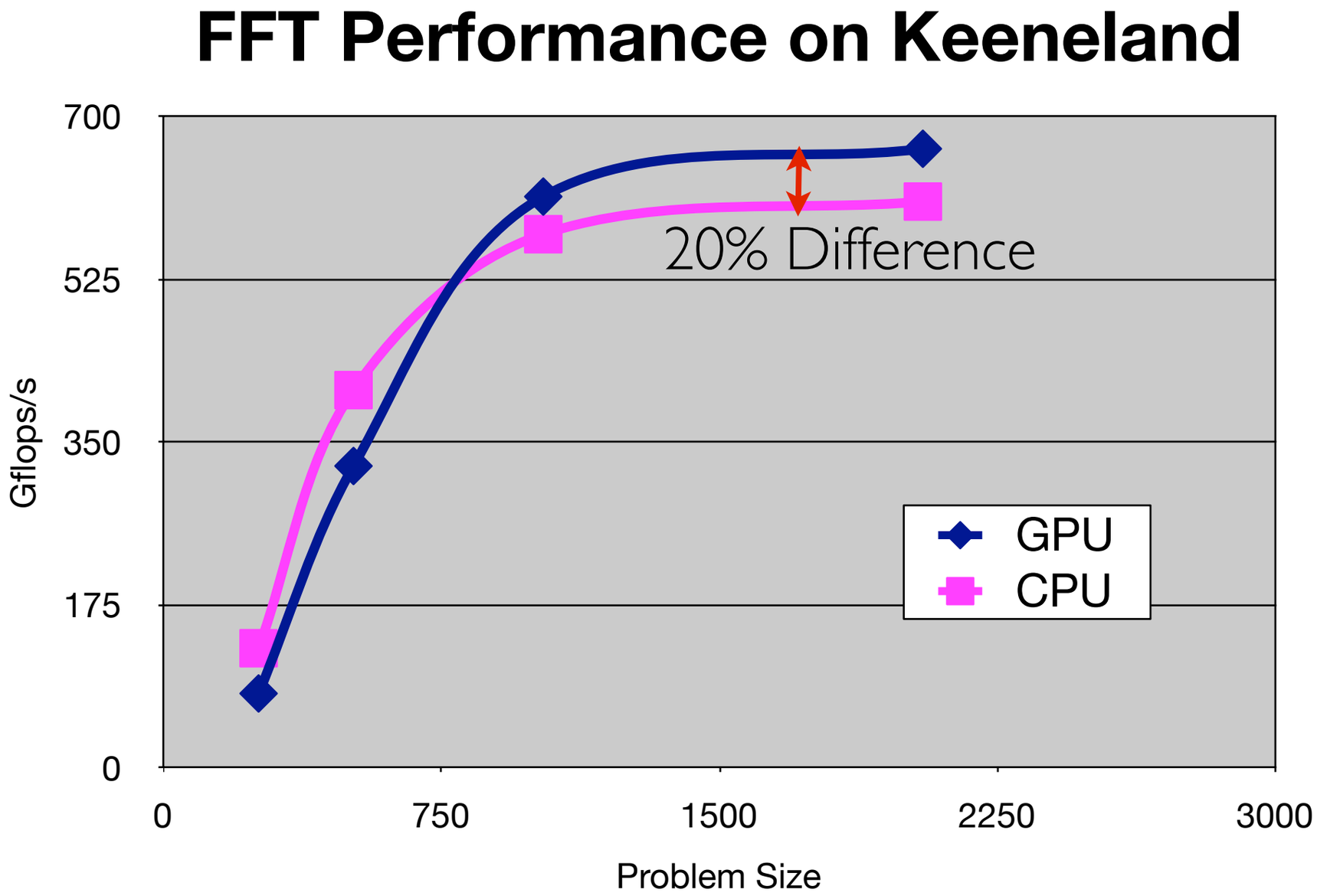
Processor Peak: 515 GF/s
Cores: 448
Memory BW: 144 GB/s
Fast Memory: 2.7 MB
Link BW: 2 GB/s

Artifacts

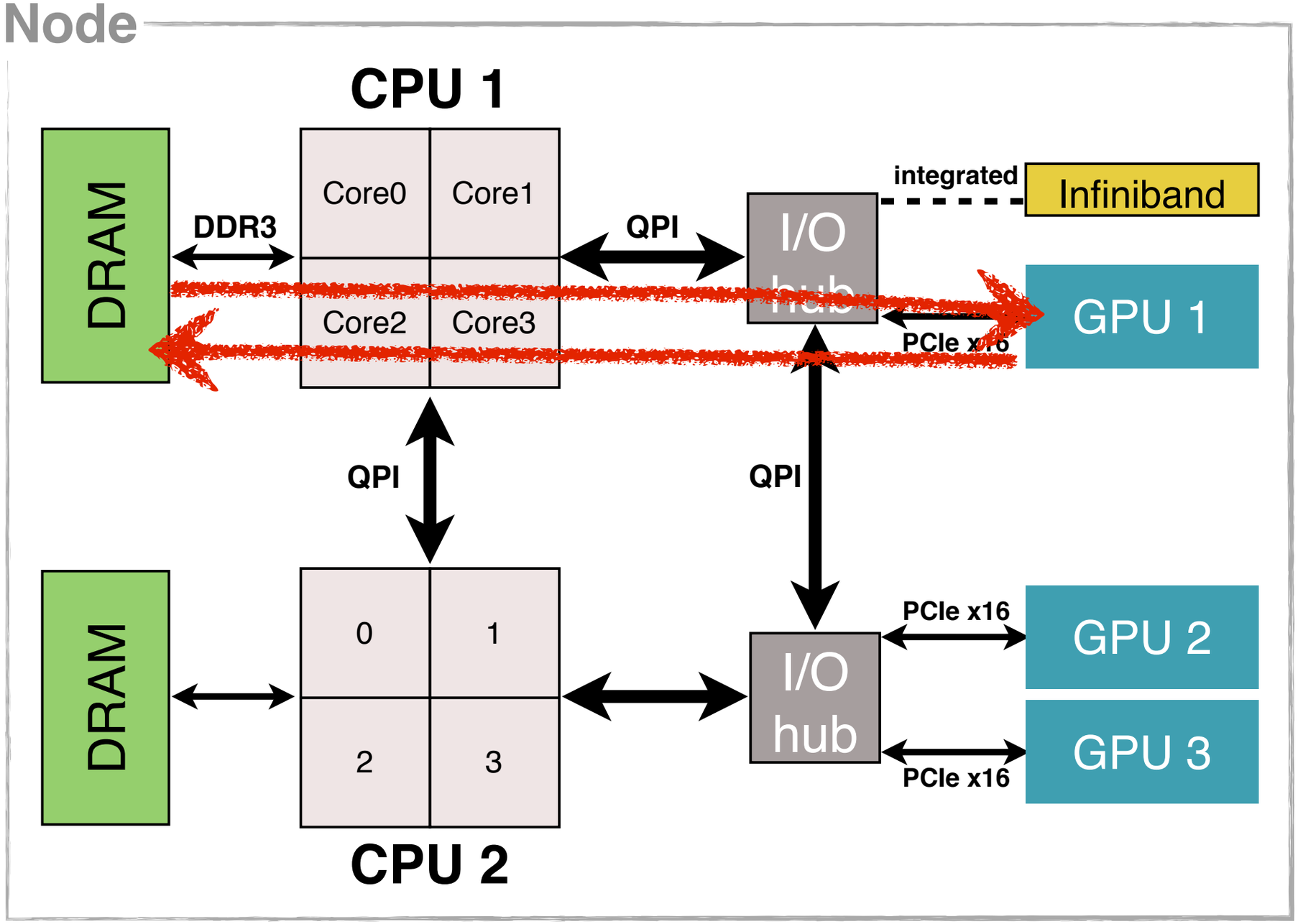
GPU vs CPU Performance



GPU vs CPU Performance

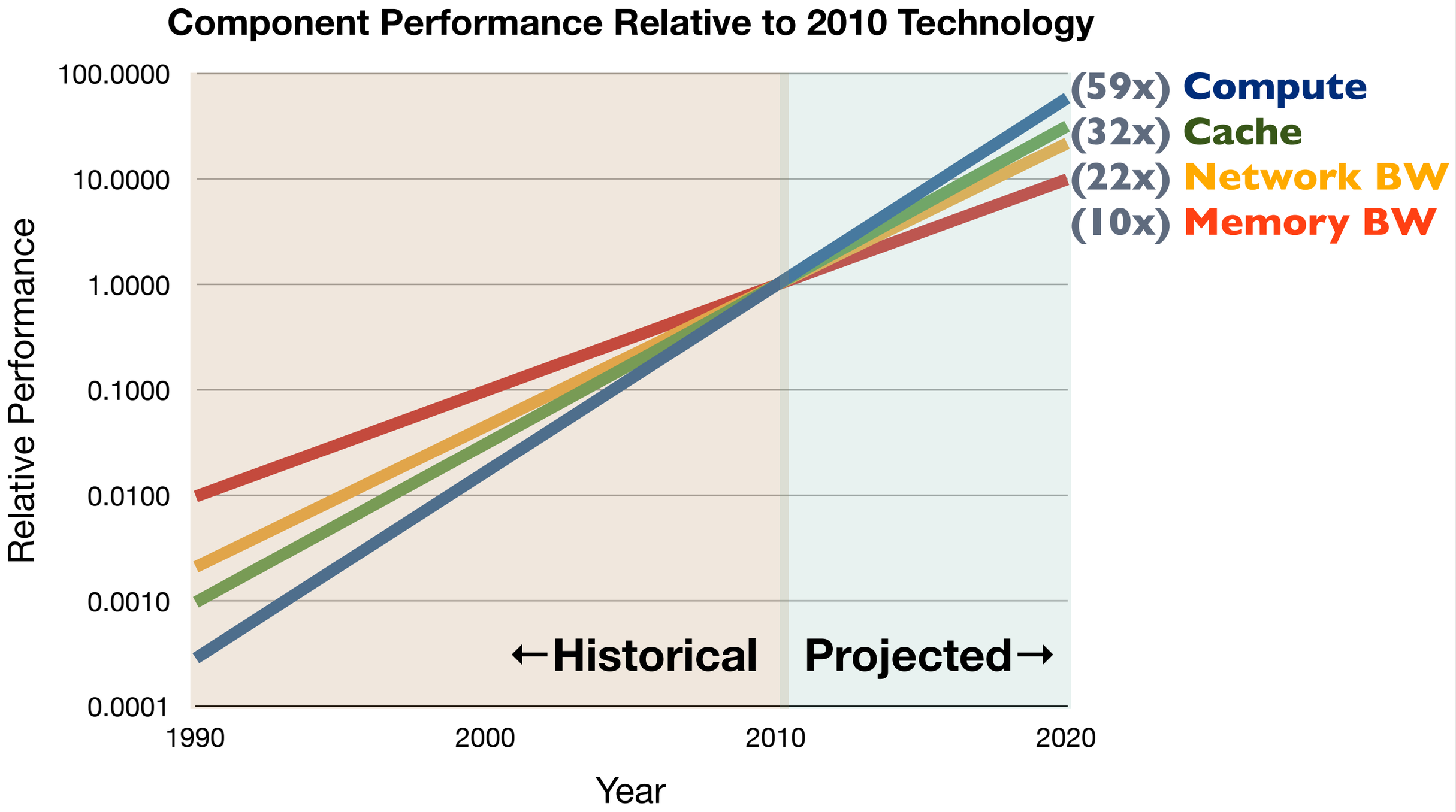


Artifacts - PCIe Bottleneck



Projections

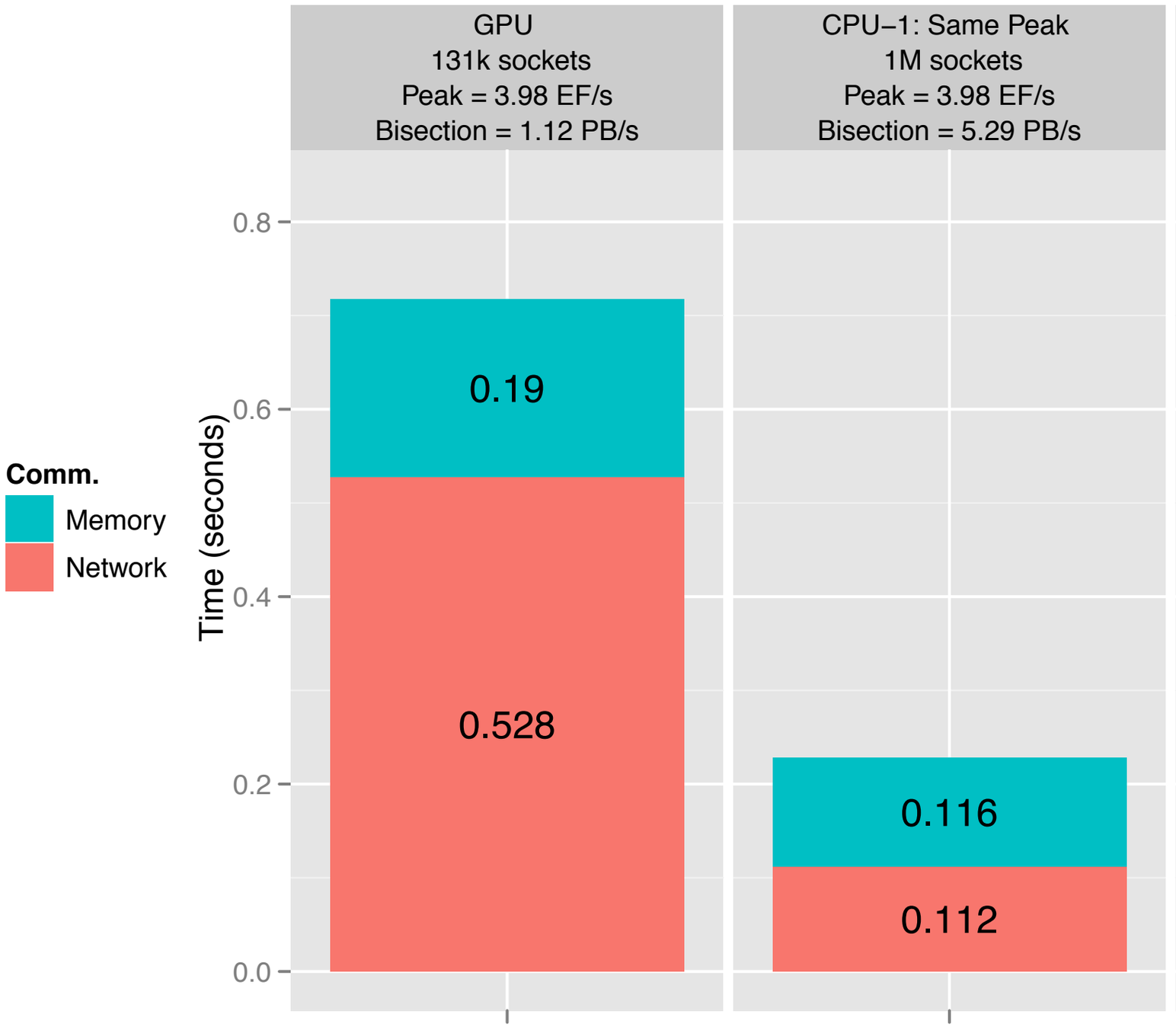
Predicting 2020 Technology



Technology Extrapolation

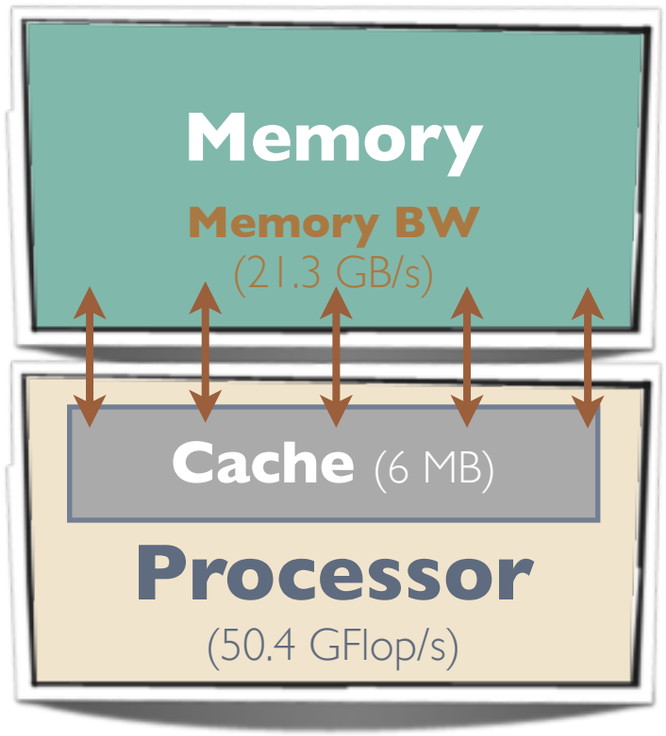
	2010	2020
	**CPU-Based** Processor Peak: 50.4 GF/s Memory BW: 21.3 GB/s Fast Memory: 6 MB Link BW: 10 GB/s 79,400 Processors	**CPU-Based** Processor Peak: 3 TF/s Memory BW: 206 GB/s Fast Memory: 192 MB Link BW: 218 GB/s 1.3 M Processors
	**GPU-Based** Processor Peak: 515 GF/s Memory BW: 144 GB/s Fast Memory: 2.7 MB Link BW: 10 GB/s 6,392 Processors	**GPU-Based** Processor Peak: 30 TF/s Memory BW: 1.4 TB/s Fast Memory: 86.4 MB Link BW: 218 GB/s 135,000 Processors

3D FFTs at Exascale (2020, n=21000)



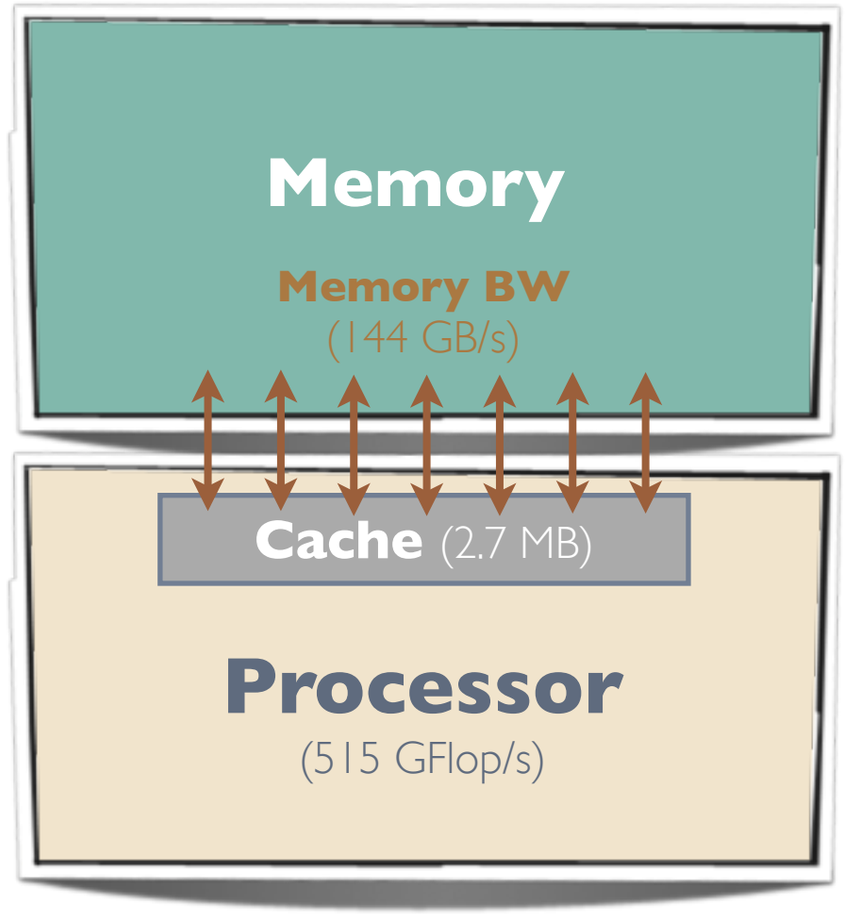
Performance vs Balance

CPU



Flop/s : Byte/s = 2.3

GPU



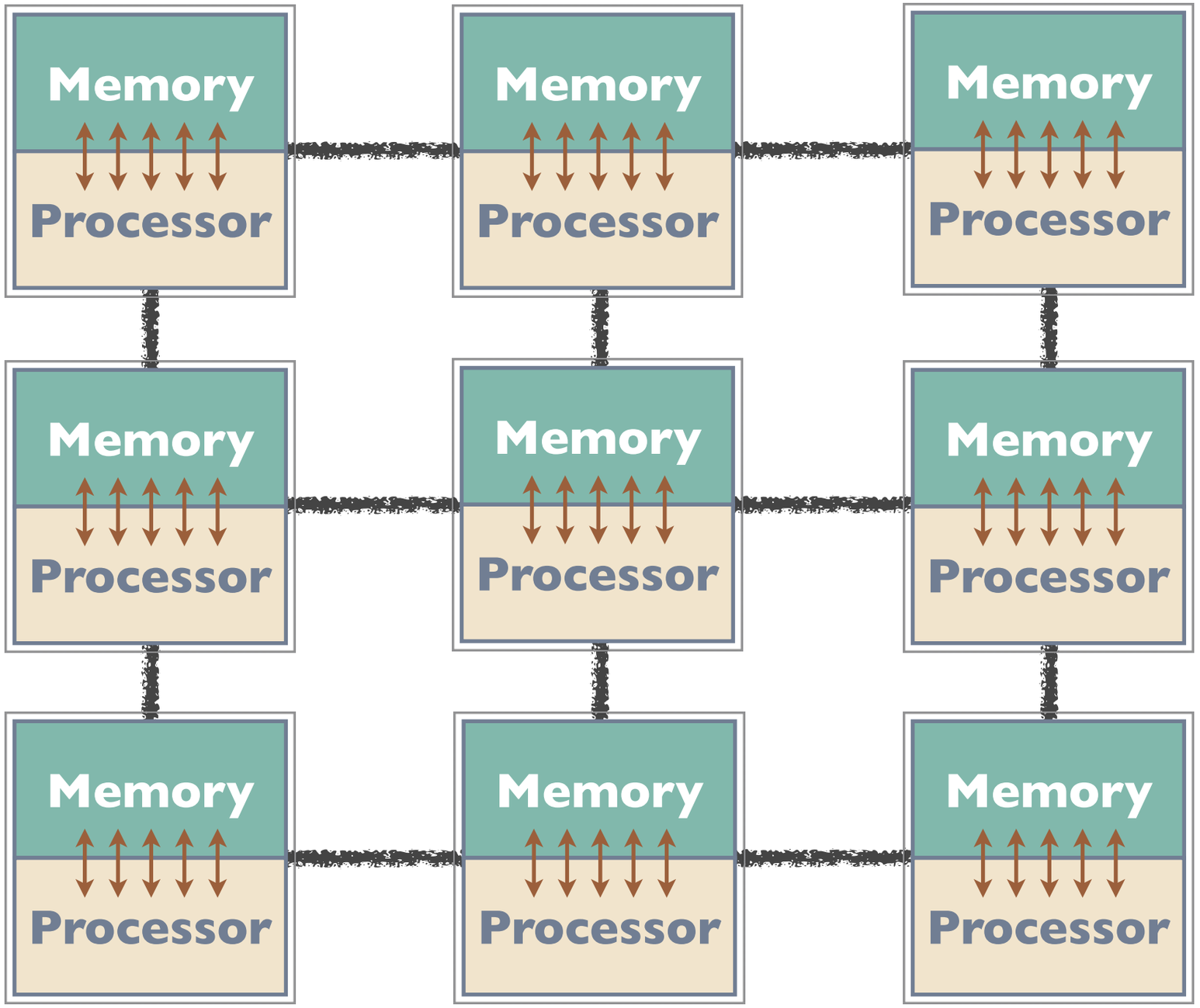
Flop/s : Byte/s = 3.6

6.7x

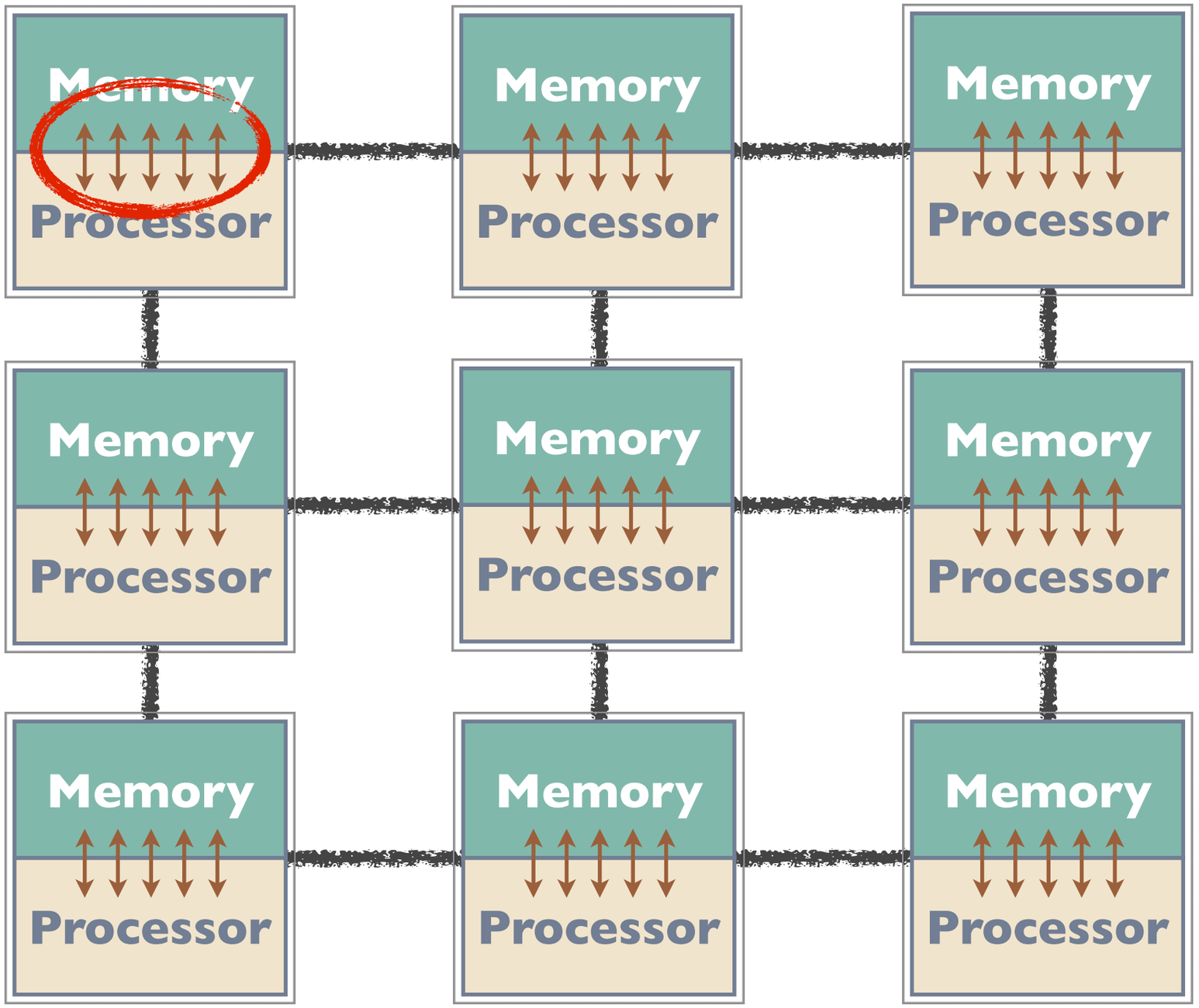
10.2x

The GPU offers better performance, but is less balanced

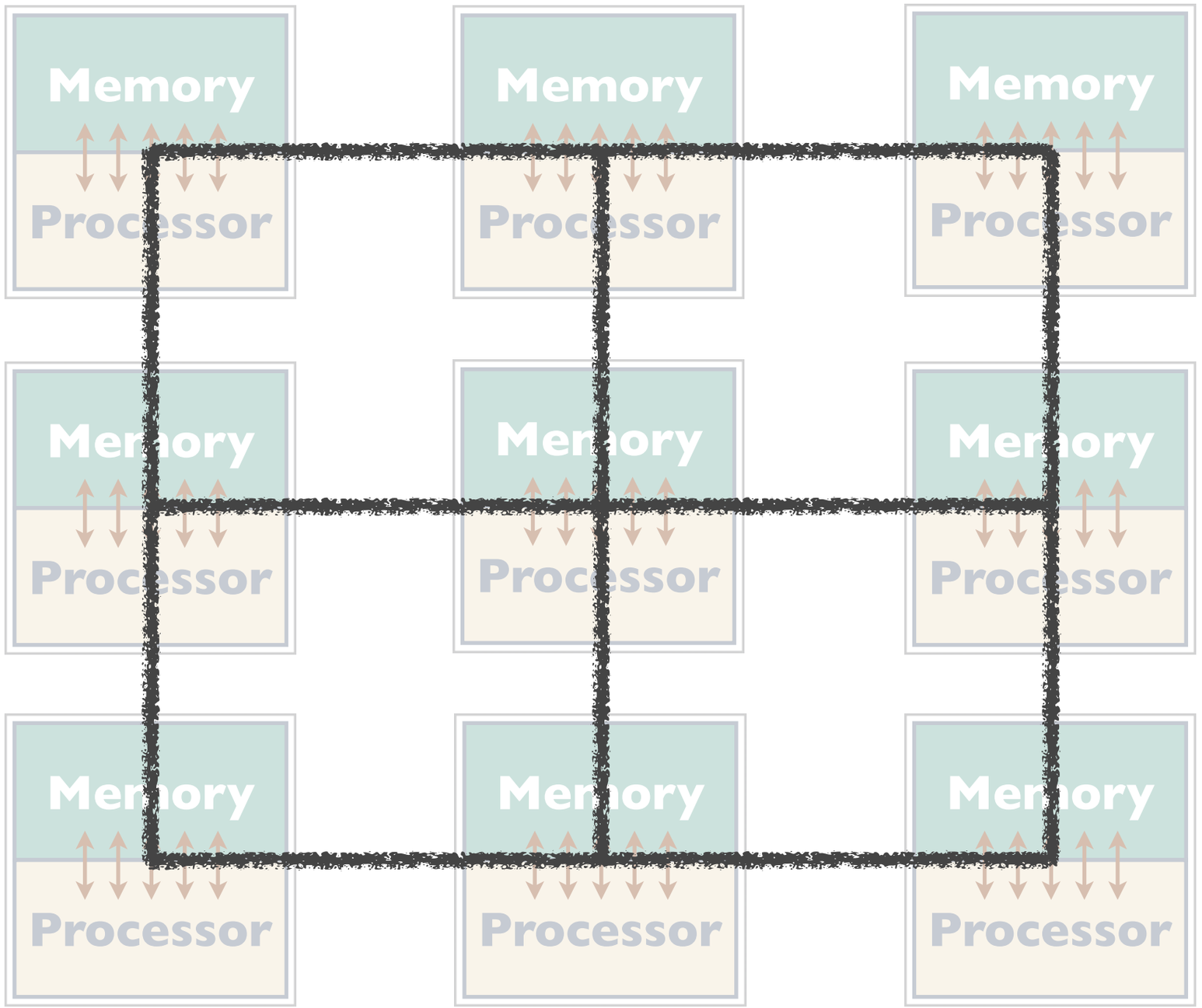
Two Costs: T_{memory} + $T_{network}$



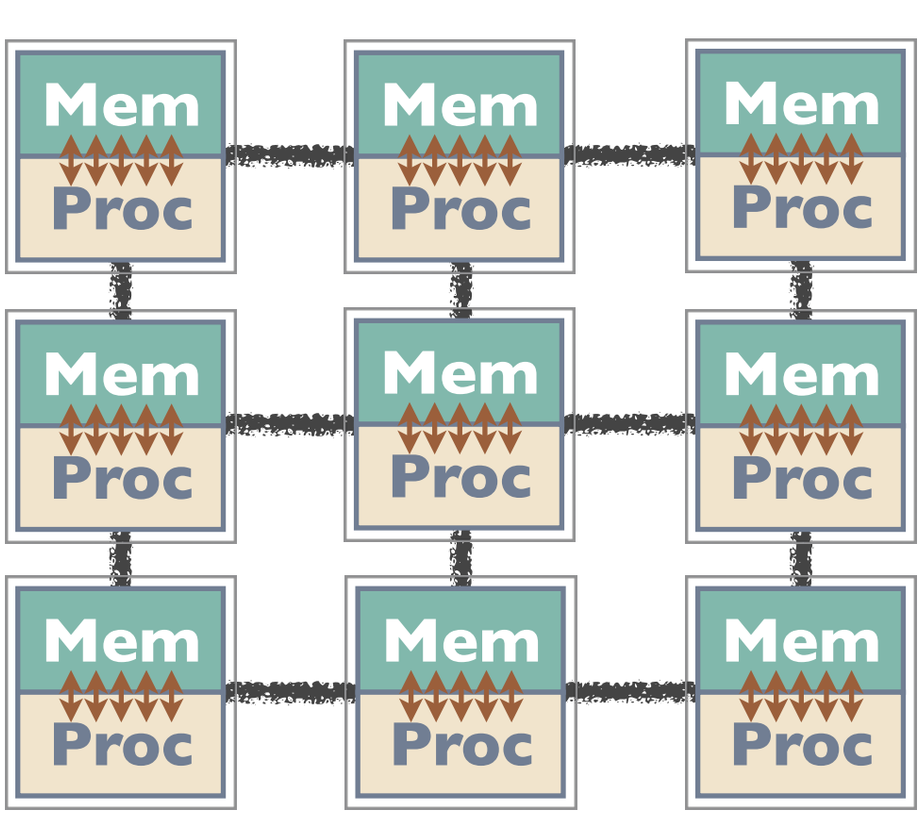
Two Costs: T_{memory} + $T_{network}$



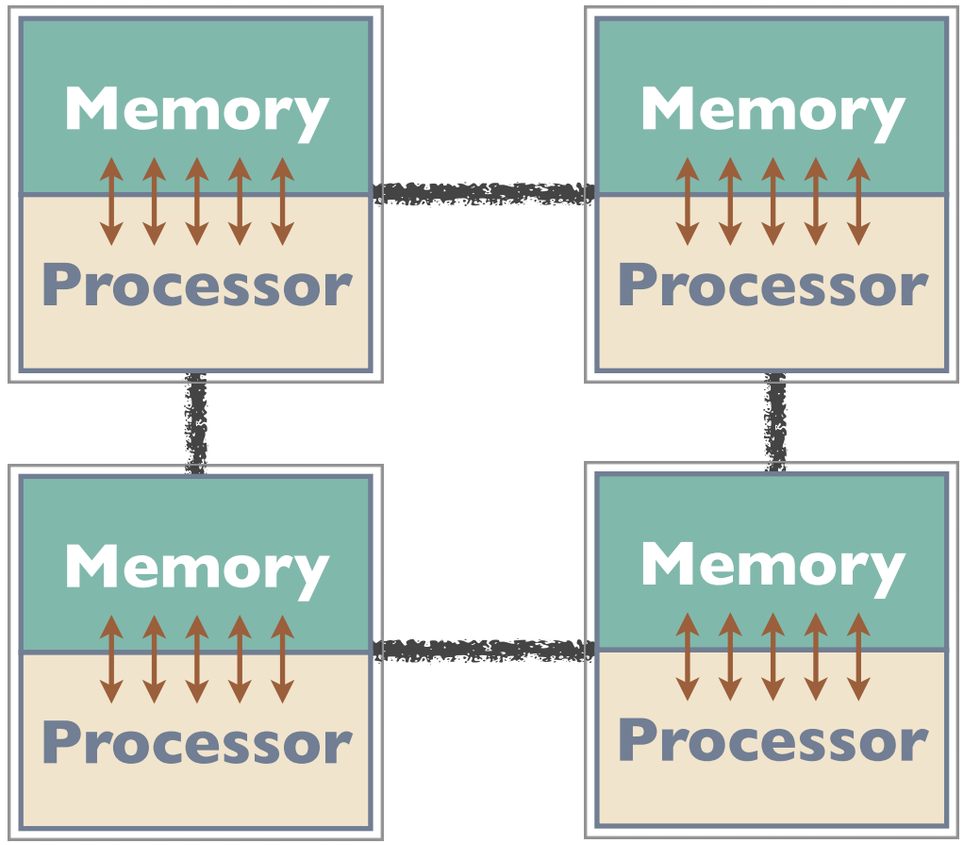
Two Costs: T_{memory} + $T_{network}$



Impact of Machine Balance



vs.

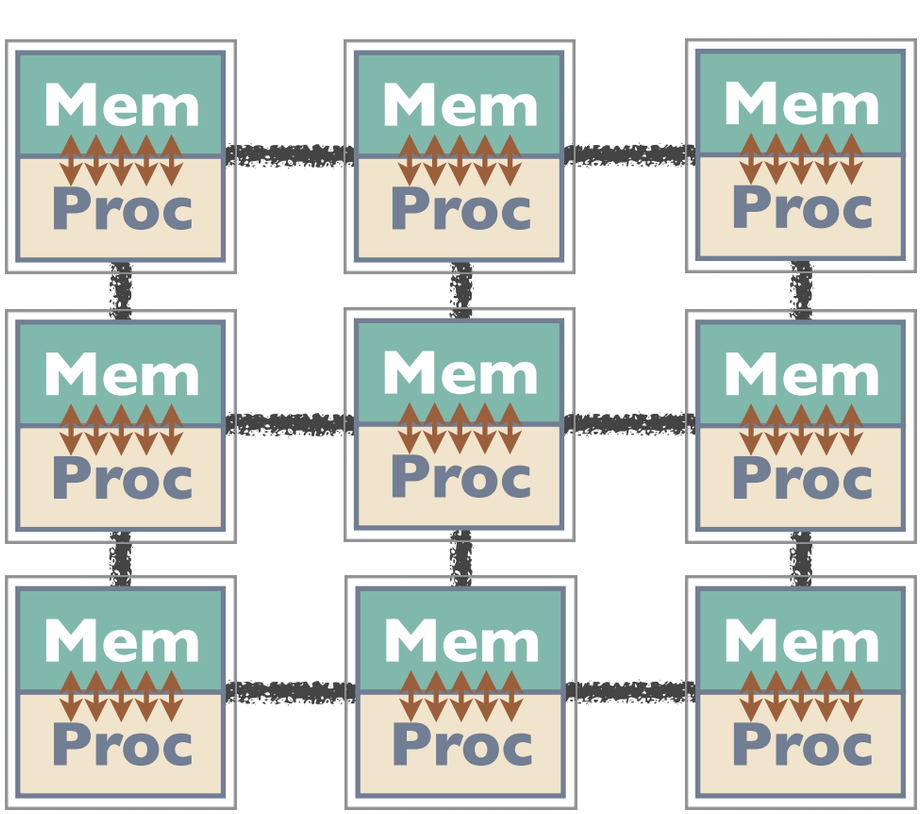


Number of processors: R_{peak} / C_{node}

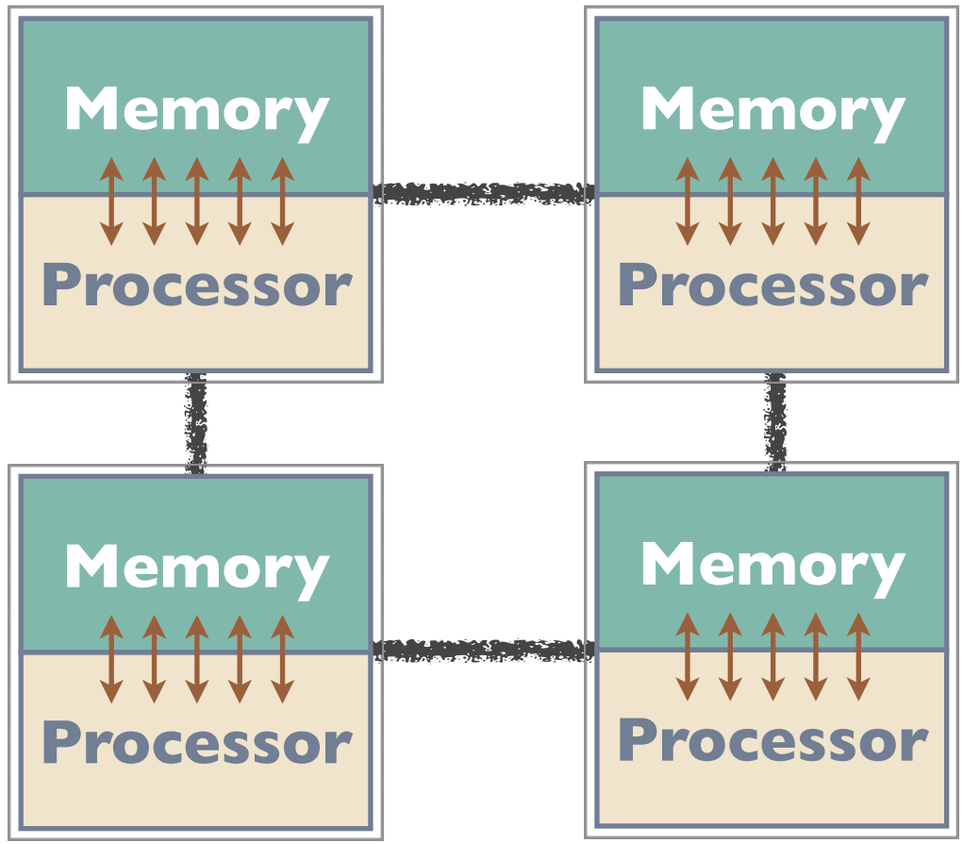
$$T_{\text{mem}} \approx \mathcal{O}\left(\frac{1}{R_{\text{peak}}} \cdot \frac{C_{\text{node}}}{\beta_{\text{mem}}} \cdot n^3 \log_Z n\right)$$

$$T_{\text{net}} \approx \mathcal{O}\left(\frac{1}{R^{\kappa}_{\text{peak}}} \cdot \frac{C^{\kappa}_{\text{node}}}{\beta_{\text{link}}} \cdot n^3\right)$$

Impact of Machine Balance



vs.

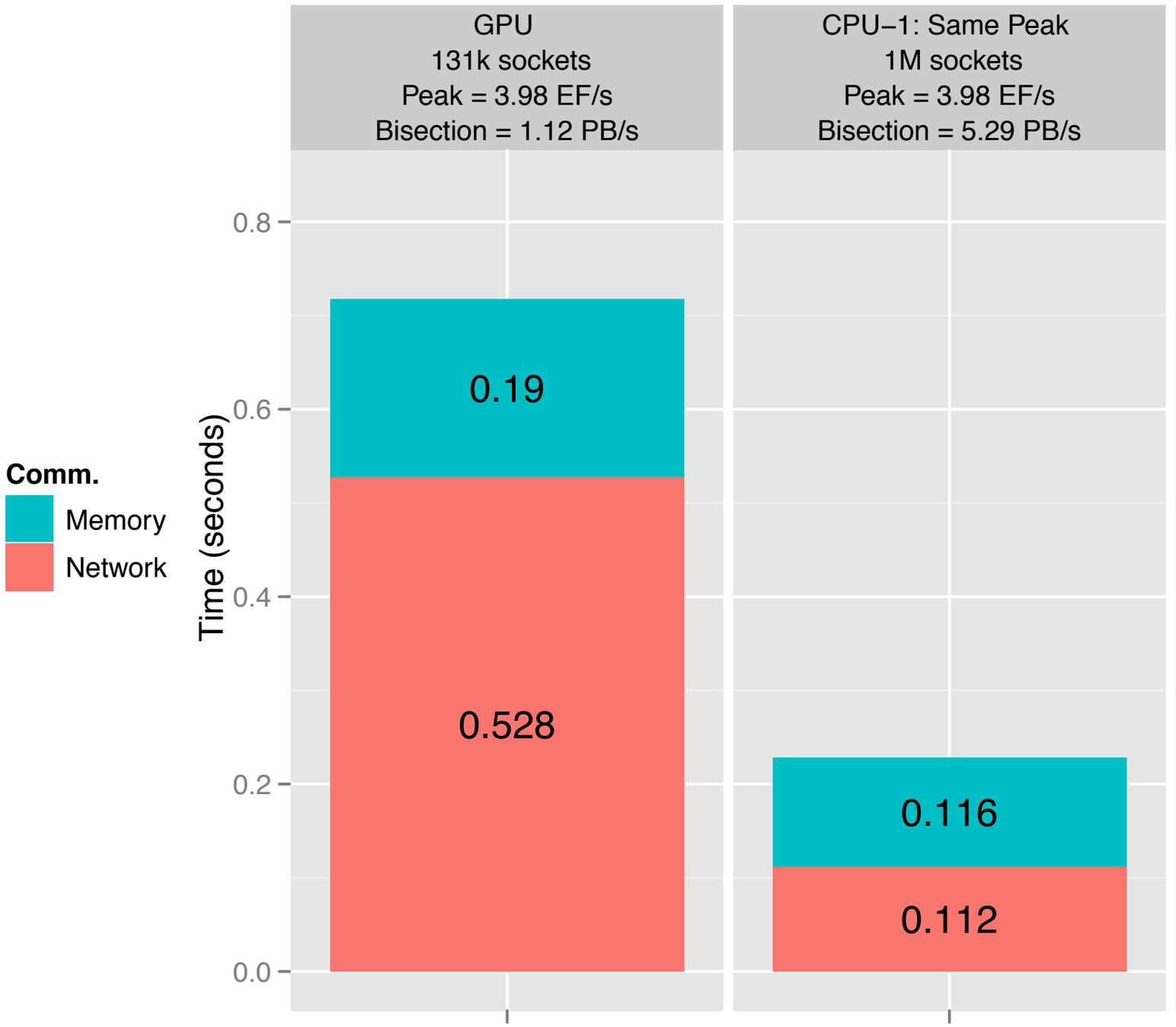


Number of processors: R_{peak} / C_{node}

$$T_{\text{mem}} \approx \mathcal{O}\left(\frac{1}{R_{\text{peak}}} \cdot \frac{C_{\text{node}}}{\beta_{\text{mem}}} \cdot n^3 \log_Z n\right)$$

$$T_{\text{net}} \approx \mathcal{O}\left(\frac{1}{R^{\kappa}_{\text{peak}}} \cdot \frac{C^{\kappa}_{\text{node}}}{\beta_{\text{link}}} \cdot n^3\right)$$

3D FFTs at Exascale (2020, n=21000)



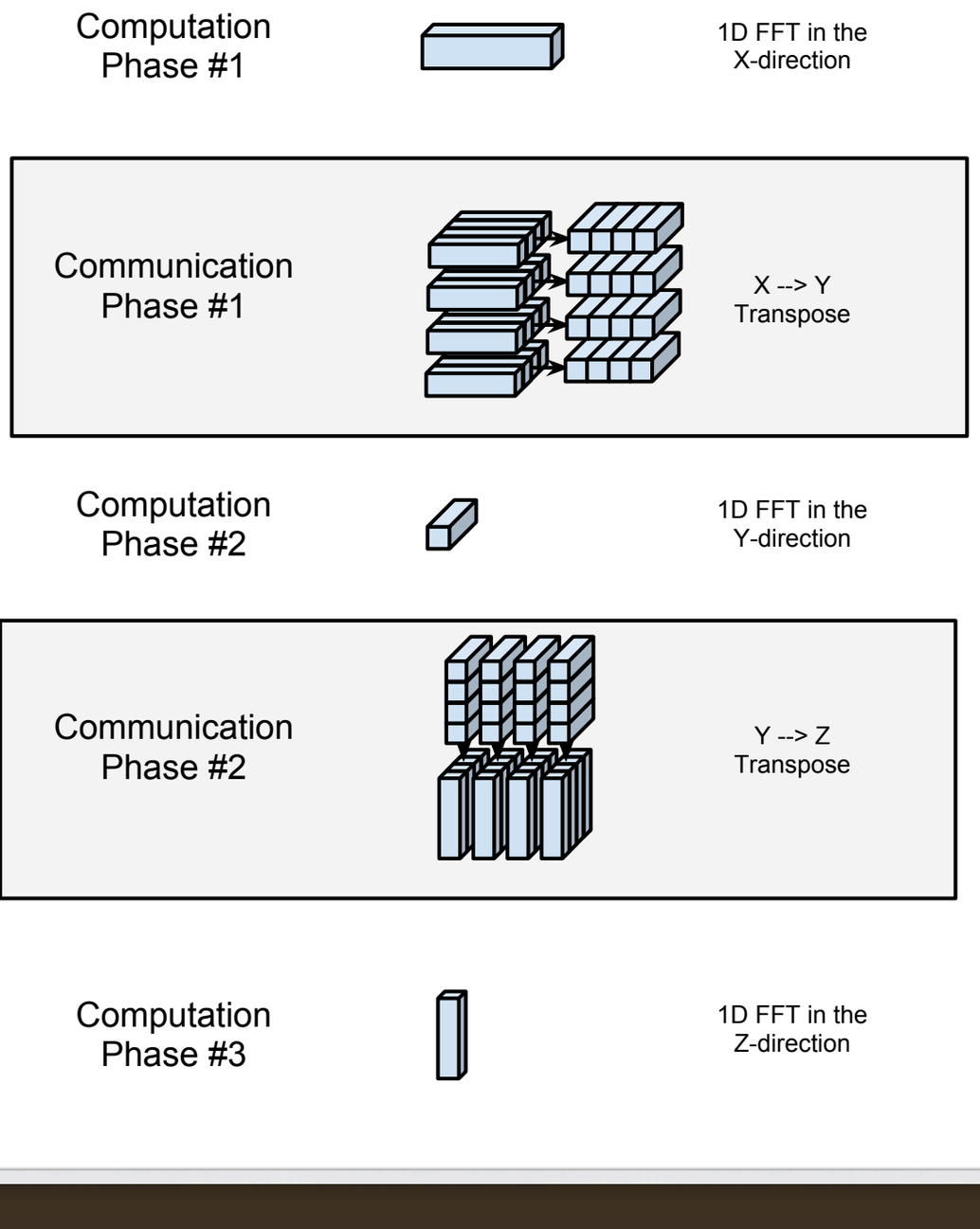
3D FFTs at Exascale (2020, n=21000)

Questions?

Parameter		2010 values	Doubling time (in years)	10-year increase factor	value
Peak:	C_{CPU} C_{GPU}	50.4 GF/s 515 GF/s	1.7	59.0×	3.0 TF/s 30 TF/s
Cores:[a]	ρ_{CPU} ρ_{GPU}	6 448	1.87	40.7×	134 18k
Memory bandwidth:	β_{CPU} β_{GPU}	21.3 GB/s 144 GB/s	3.0	9.7×	206 GB/s 1.4 TB/s
Fast memory	Z_{CPU} Z_{GPU}	6 MB 2.7 MB[b]	2.0	32.0×	192 MB 86.4 MB
Line size:	L_{CPU} L_{GPU}	64 B 128 B	10.2	2.0×	128 B 256 B
Link bandwidth:	β_{link}	10 GB/s	2.25	21.8×	218 GB/s
Machine peak:	R_{peak}	4 PF/s	1.0	1000×	4 EF/s
System memory:	E	635 TB	1.3	208×	132 PB
Nodes ($\frac{R_{\text{peak}}}{C}$):	P_{CPU} P_{GPU}	79,400 7,770	2.4	17.4×	1.3M 135,000

Distributed 3D FFT - Performance Model

3D FFT Using the Pencil Decomposition of the Transpose Method

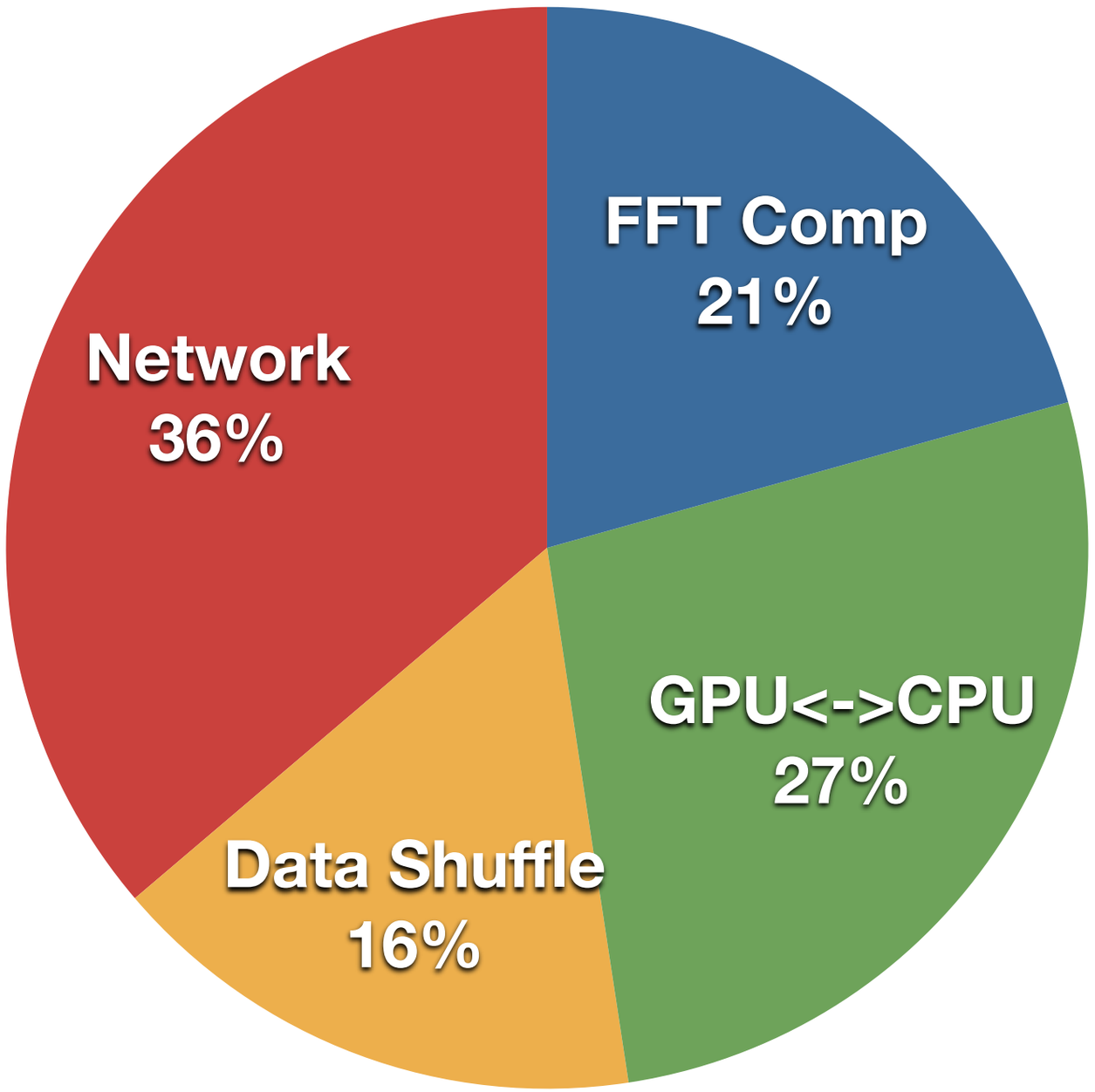


$$T_{\text{mem}} \approx 3 \times \frac{n^2}{P} \cdot \frac{A \times n(\max(\log_Z n, 1.0))}{\beta_{\text{mem}}}$$

$$T_{\text{net}} \approx 2 \times \frac{n^3}{P^{\frac{2}{3}} \beta_{\text{link}}}$$

Cache Capacity (Z)
Nodes (P)
Memory BW ($\boldsymbol{\beta}_{\text{mem}}$)
Network BW($\boldsymbol{\beta}_{\text{net}}$)

3D FFT on GPU Cluster



FFT Performance on Hopper