Understanding The Effects of MPI Oversubscription & Multi-Process Service on Distributed-memory FFT LIbraries

Daniel Schultz

Friday Feb 21, 2020

Data Distribution

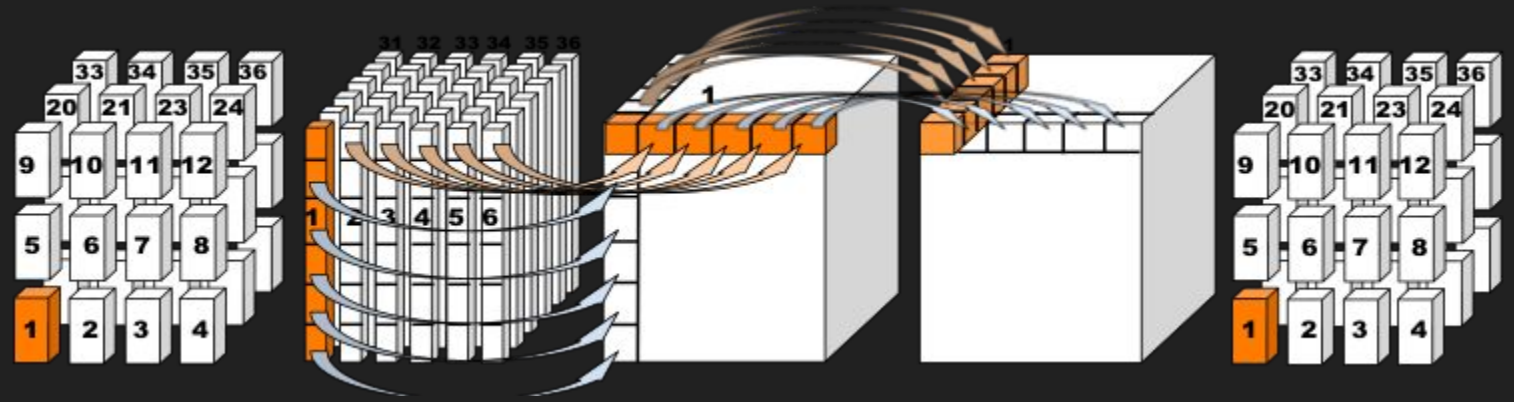
Pencil Based Libraries

- heFFTe - 4 step data distribution
- FFTMPI - 4 step data distribution

Cube Based Libraries

- SWFFT (ANL) - 6 step data distribution

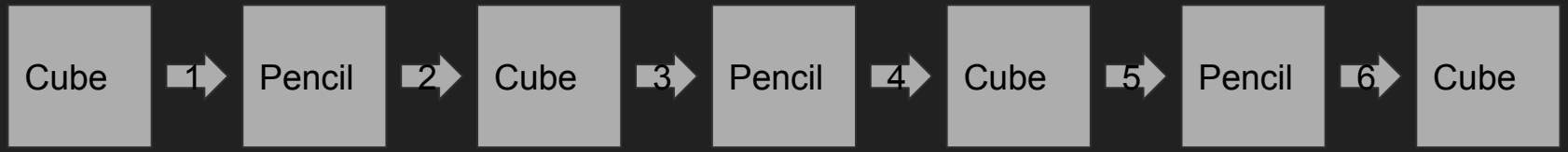
heFFTe Data Distribution



- Pencil -> Pencil
- 4 steps

- FLOPS: $5 * n^3 * \log(n^3)$
- Data IO: $16 * 4 * n^3$ (DP Complex)
- CI: $15 * \log(n^3) / 64$

SWFFT Data Distribution



- Cube -> Cube
- 6 steps

- FLOPS: $5 * n^3 * \log(n^3)$
- Data IO: $16 * 6 * n^3$ (DP Complex)
- CI: $15 * \log(n^3) / 96$

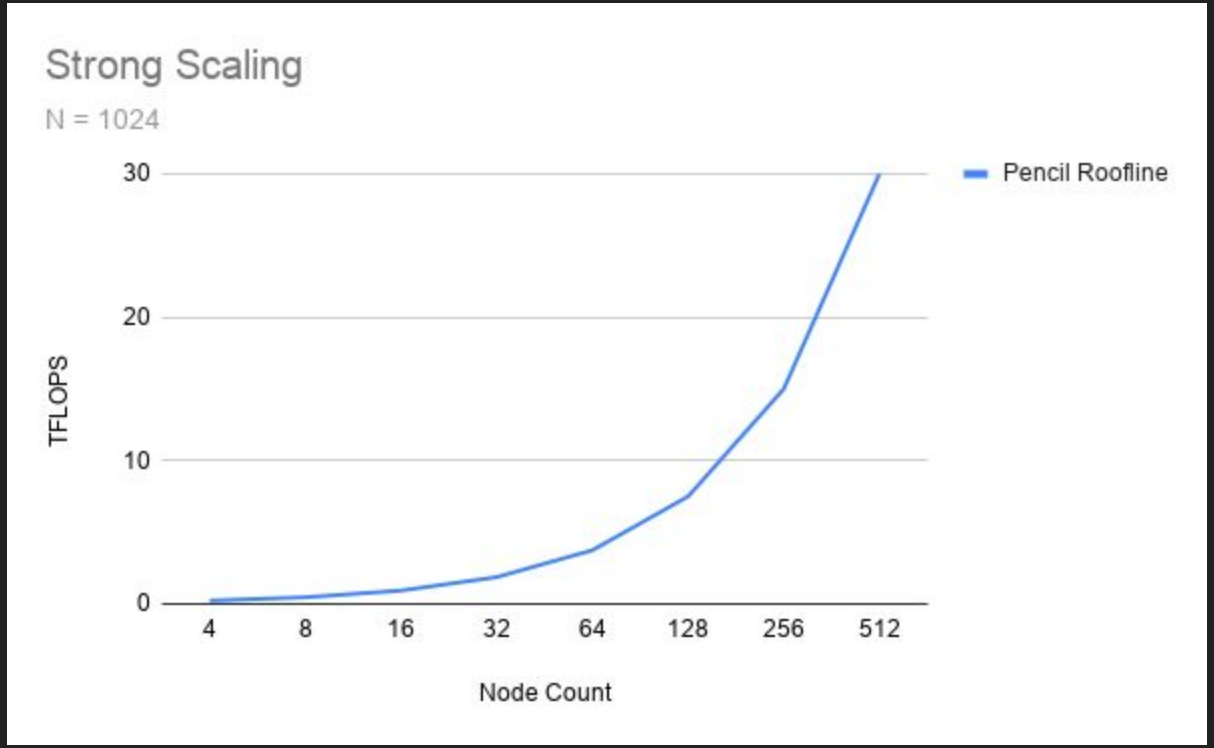
Comparison

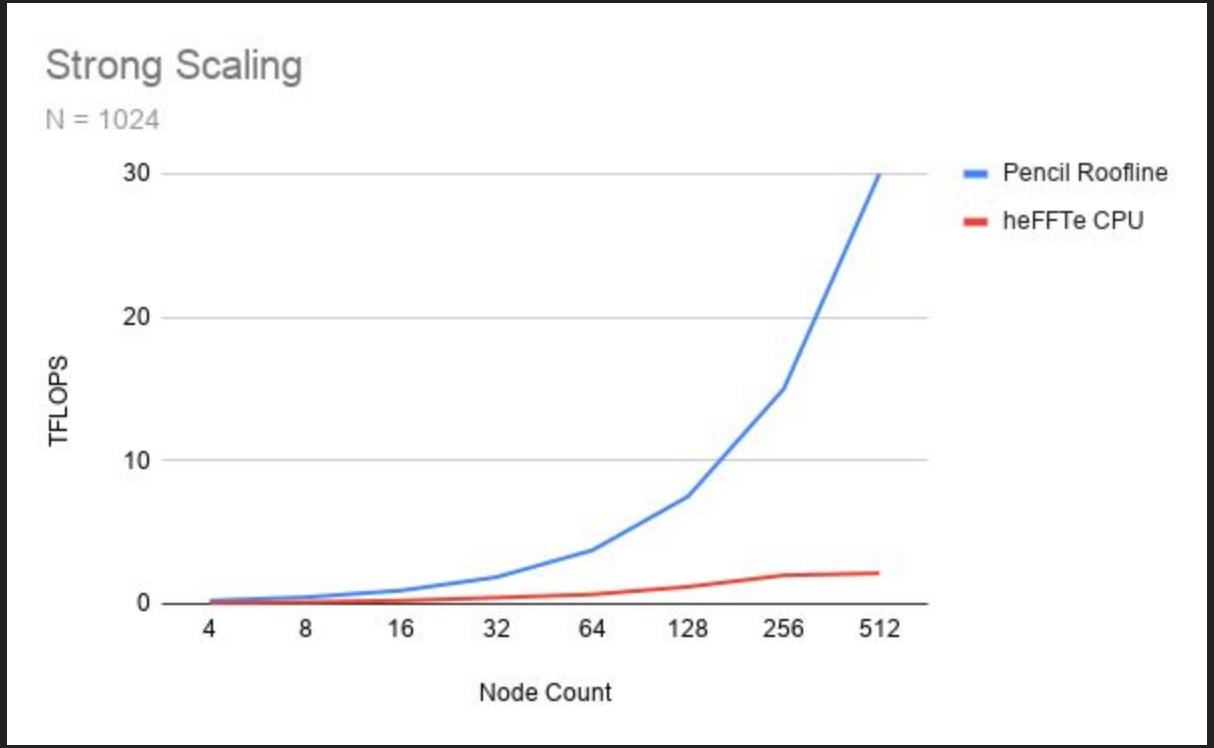
Pencils Vs Cubes

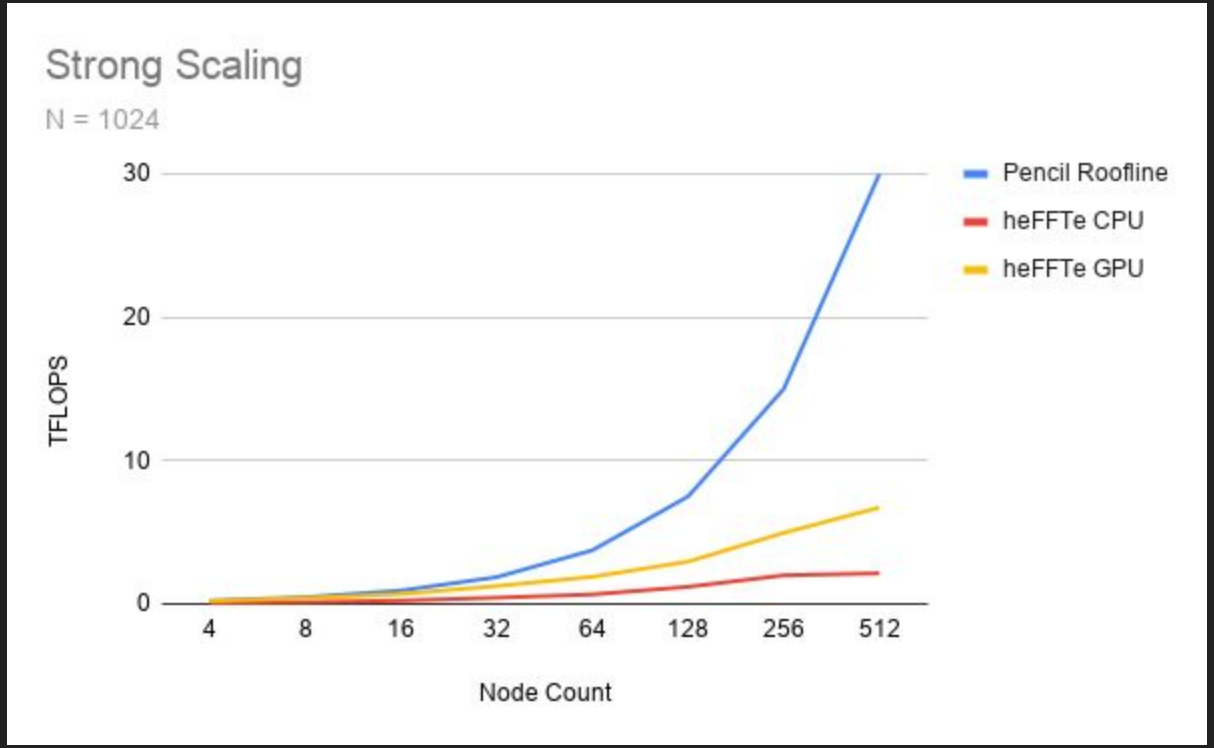
- Same flops
- 1.5x data IO on cubes
- Better AI on pencils as a result

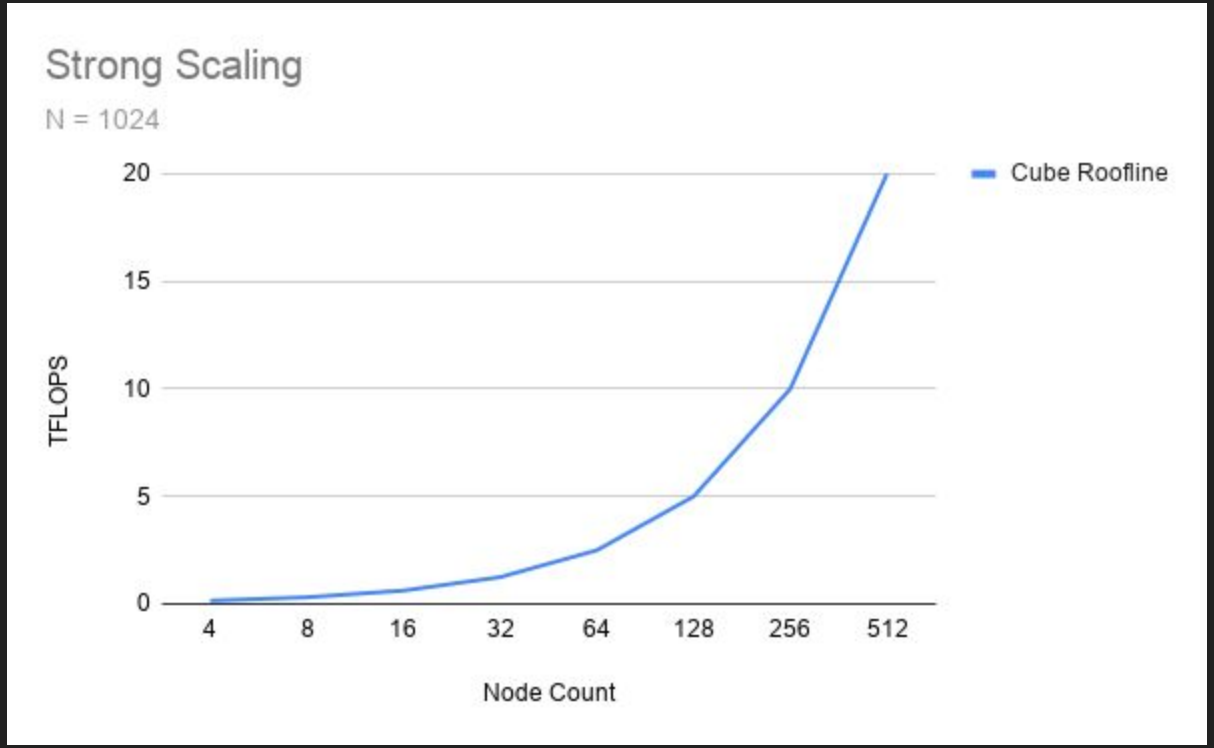
Pencils AI: 2.34

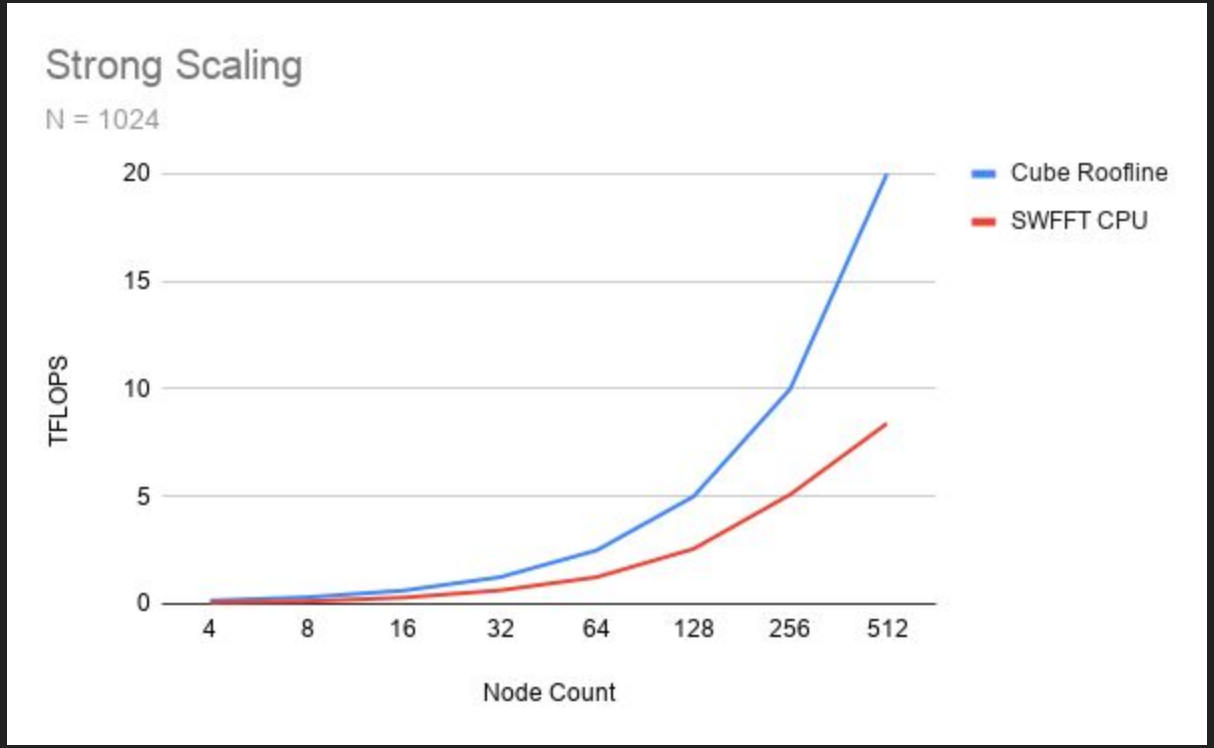
Cubes AI: 1.56

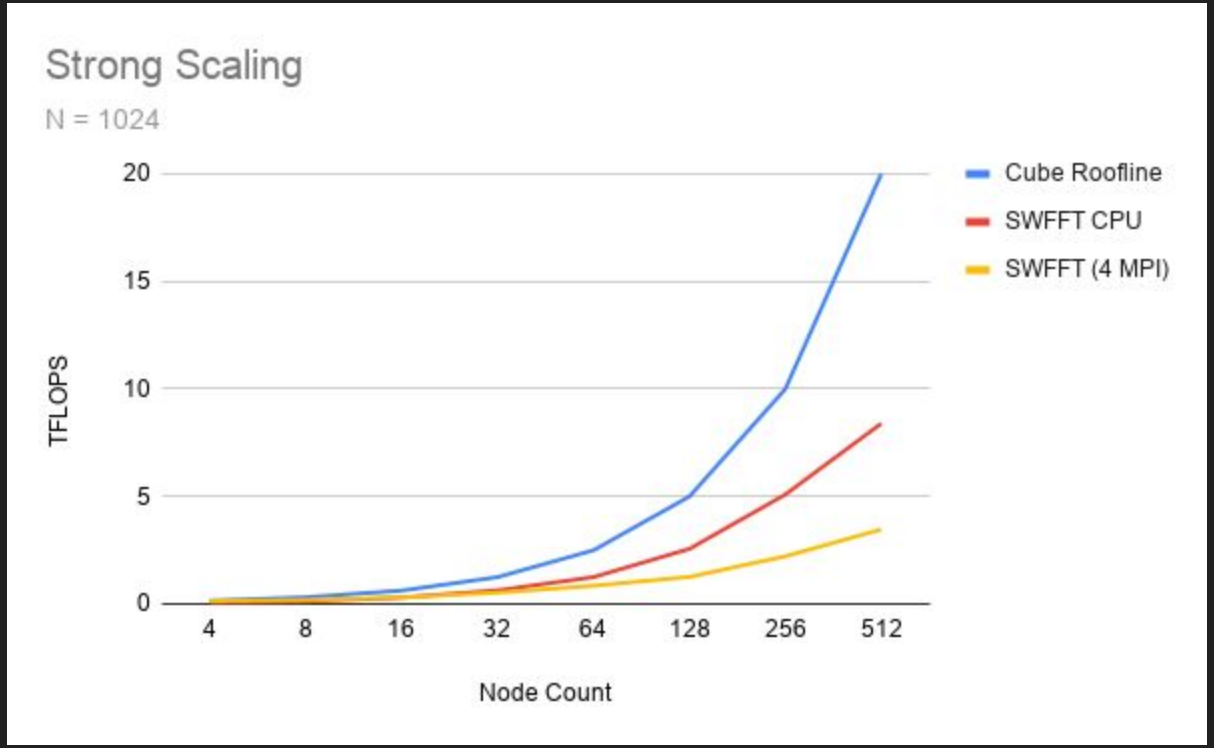

Strong Scaling
N = 1024
30
20
10
0
TFLOPS
4
8
16
32
64
128
256
512
Node Count
Pencil Roofline


Strong Scaling
N = 1024
30
20
10
0
TFLOPS
4
8
16
32
64
128
256
512
Node Count
Pencil Roofline
heFFTe CPU


Strong Scaling
N = 1024
30
20
10
0
TFLOPS
4
8
16
32
64
128
256
512
Node Count
Pencil Roofline
heFFTe CPU
heFFTe GPU


Strong Scaling
N = 1024
20
15
10
5
0
TFLOPS
4
8
16
32
64
128
256
512
Node Count
Cube Roofline


Strong Scaling
N = 1024
20
15
10
5
0
TFLOPS
4
8
16
32
64
128
256
512
Node Count
Cube Roofline
SWFFT CPU


Strong Scaling
N = 1024
20
15
10
5
0
TFLOPS
4
8
16
32
64
128
256
512
Node Count
Cube Roofline
SWFFT CPU
SWFFT (4 MPI)

Is bandwidth a issue?

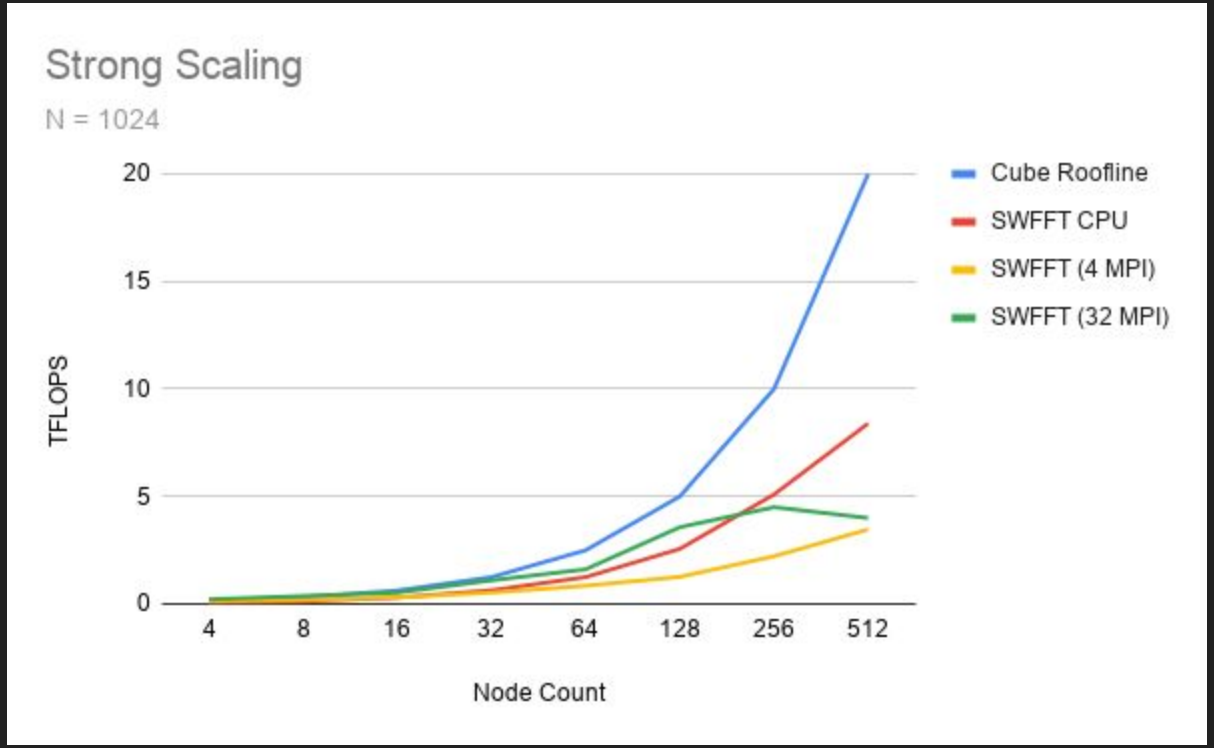
GPU Multi-process Service

MPI

- Normal: One MPI Rank -> One CPU Core “slot”
- Oversubscribed: Multiple MPI Ranks -> One CPU Core

Using GPUs with MPI

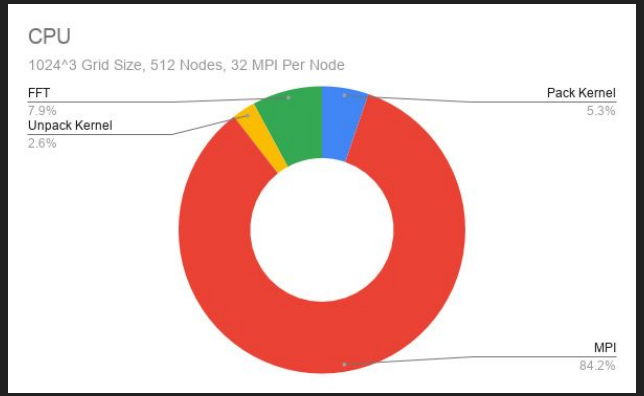
- Normal: One MPI Rank -> One GPU
  - One Kernel -> One Rank -> One GPU (What happens if the kernel is underutilizing?)
- Oversubscribed: Multiple MPI Ranks -> One GPU
  - Many Kernels -> Many Ranks -> One GPU (Can be multiple invocations of the same kernel.)

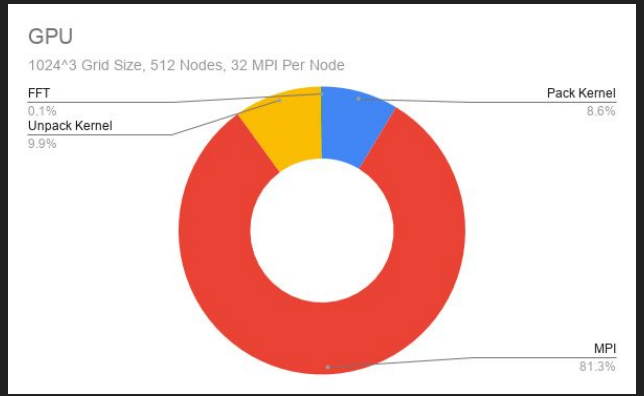

Strong Scaling
N = 1024
20
15
10
5
0
TFLOPS
4
8
16
32
64
128
256
512
Node Count
Cube Roofline
SWFFT CPU
SWFFT (4 MPI)
SWFFT (32 MPI)

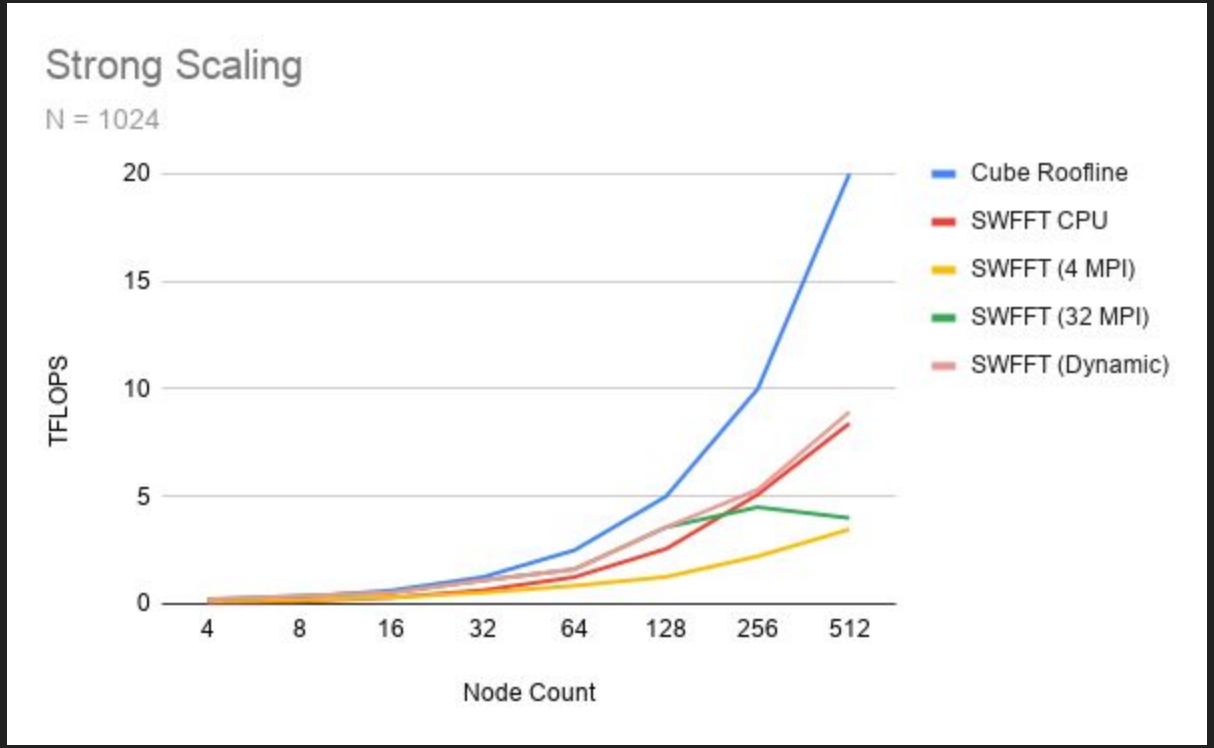
	GFLOPS	Running Time	Pack Kernel	MPI	Unpack Kernel	FFT
CPU	8394.752	0.019	0.001	0.016	0.0005	0.0015
GPU	3991.06	0.0405	0.0035	0.032945	0.004	0.000055

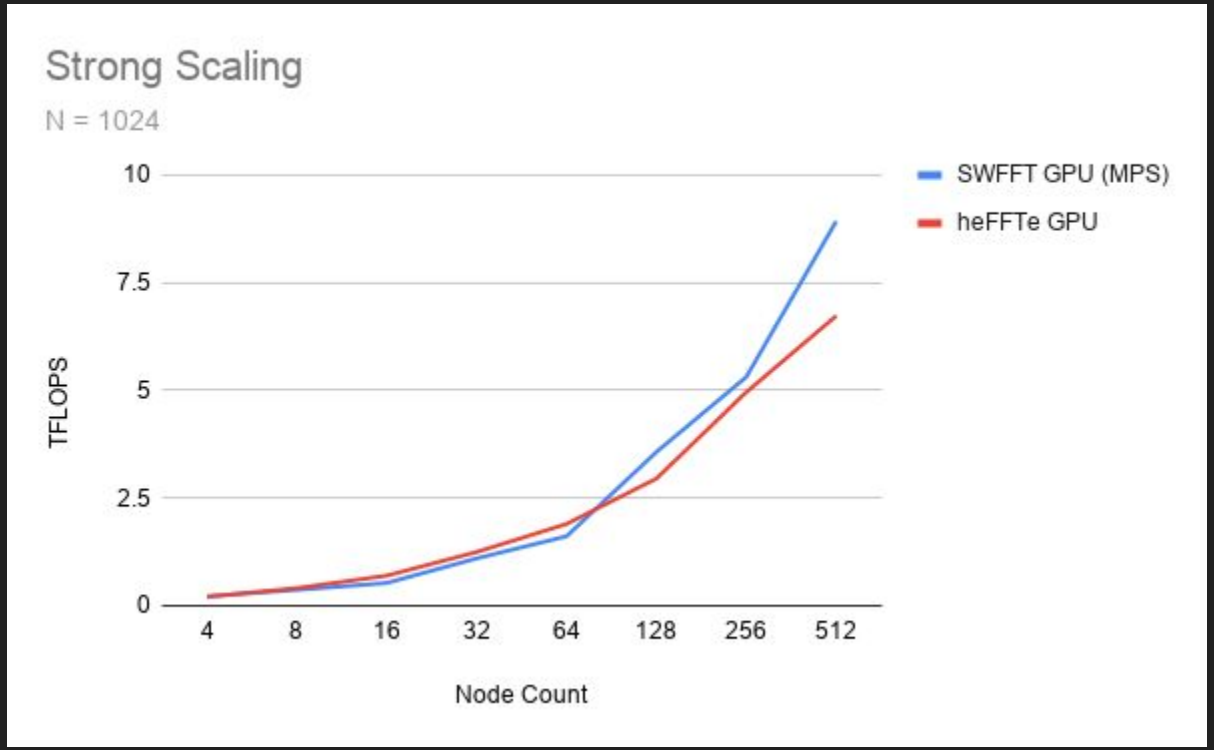
CPU -> GPU Slowdown

- Packing Kernel: 3.5x slower
- Unpacking Kernel: 8x slower
- MPI P2P: 2.06x slower
- FFT: 27x faster






Strong Scaling
N = 1024
20
15
10
5
0
TFLOPS
4
8
16
32
64
128
256
512
Node Count
Cube Roofline
SWFFT CPU
SWFFT (4 MPI)
SWFFT (32 MPI)
SWFFT (Dynamic)



Cubes Vs Pencils

Why?

Sending Pencils

- N = 1024
- Data: 16 * 4 * 1024^3 = 64 GB
- Node Count: 512
- MPI Ranks: 6 per node. 1 per GPU
- Messages per node: 512 nodes * 6 rank * 4 data distributions = 12,288 messages
- Total Messages: All-to-All = Approx 6.3 million messages
- Message Size: 10.67 KB

6.7 TFLOPS

Sending Cubes

- N = 1024
- Data: 16 * 6 * 1024^3 = 96 GB
- Node Count: 512
- MPI Ranks: 8 per node. Oversubscribed
- Messages per node: 512 nodes * 8 rank * 6 data distributions = 24576 messages
- Total Messages: Approx 12.6 million
- Message Size: 8 KB

8.9 TFLOPS