EXASCALE COMPUTING

Kate Clark @ Lattice 2023

EXASCALE IS FINALLY HERE

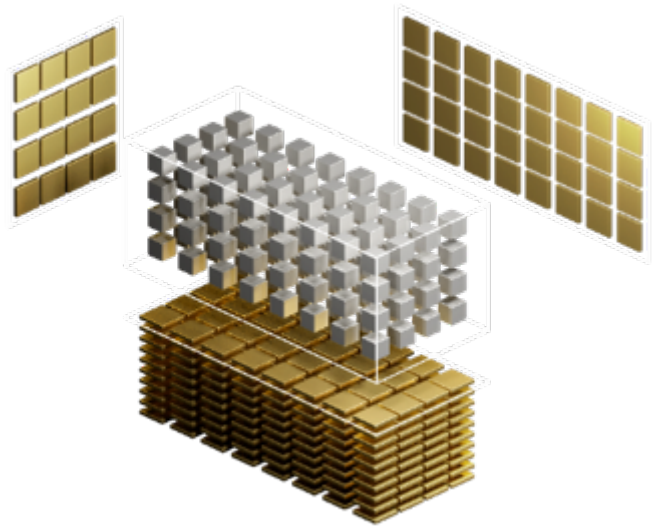
(Zettascale will be *much* harder)

Matrix and tensor operations required to saturate the machine

Low precision is much faster

Extreme parallelism required

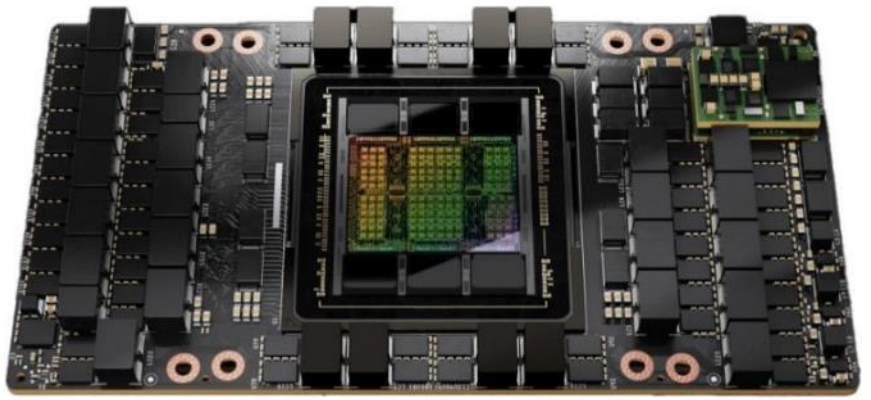
Hierarchy and Locality must be considered



A100 TO H100

Simulation Flops, AI Flops & Memory Bandwidth

Operation	Nvidia H100	Vs. A100	
FP64	34 Tflops	3x	Simulation Flops
FP64 TC	67 Tflops	3x	
FP32	67 Tflops	3x	
TF32 TC	494 Tflops	3x	AI Flops
FP16 TC	989 Tflops	3x	
INT8 TC	1979 Tera ops	3x	
FP8 TC	1979 Tflops	6x	
Memory Bandwidth	3.35 TB/sec	2x	Memory Bandwidth
External Bandwidth	1.9 TB/sec	3X	Extremal Bandwidth

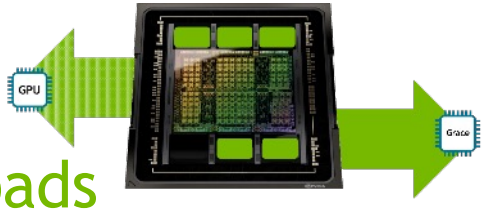


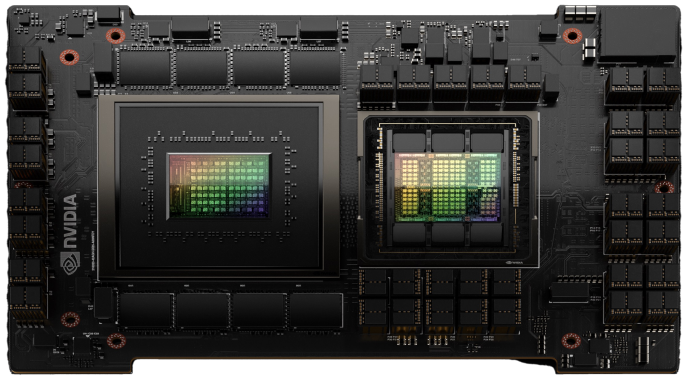
Custom TSMC 4N Process | 4.9 TB/s Total External B/W

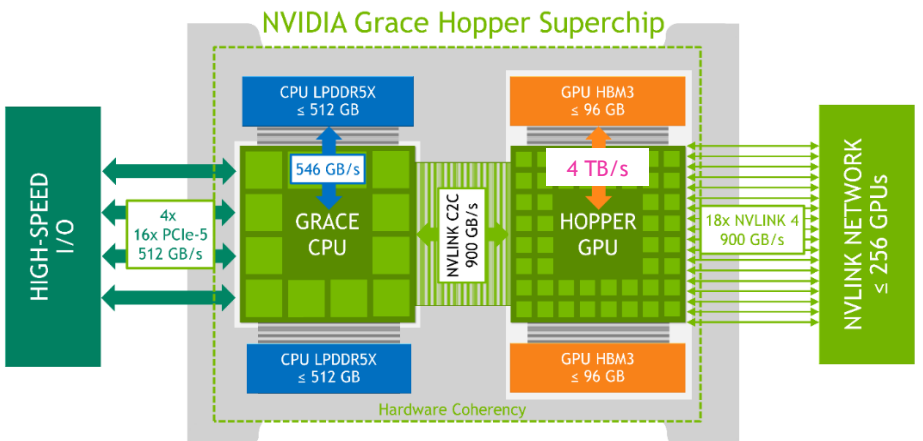
	A100	H100	
Shared Memory per Block	160 kB	228 kB	1.43X
L2 Cache Size	40 MB	50MB	1.25X

GRACE HOPPER SUPERCHIP

Supercharge performance for AI and HPC workloads







GRACE HOPPER SUPERCHIP	
GPU	Hopper 96GB HBM3, 4 TB/s, 7 MIG
CPU	72 Core ARMv9 (10 Core per MIG)
CPU Mem	LPDDR5 512GB, 4x lower power than DDR5
CPU to GPU NVLink C2C	450GB/s per dir. & cache coherent, 7x PCIe Gen5 x16
GPU to GPU NVLink MN-NVLink	450GB/s per dir., up to 256 GPUs per NVLink Domain
TDP	Max 1kW

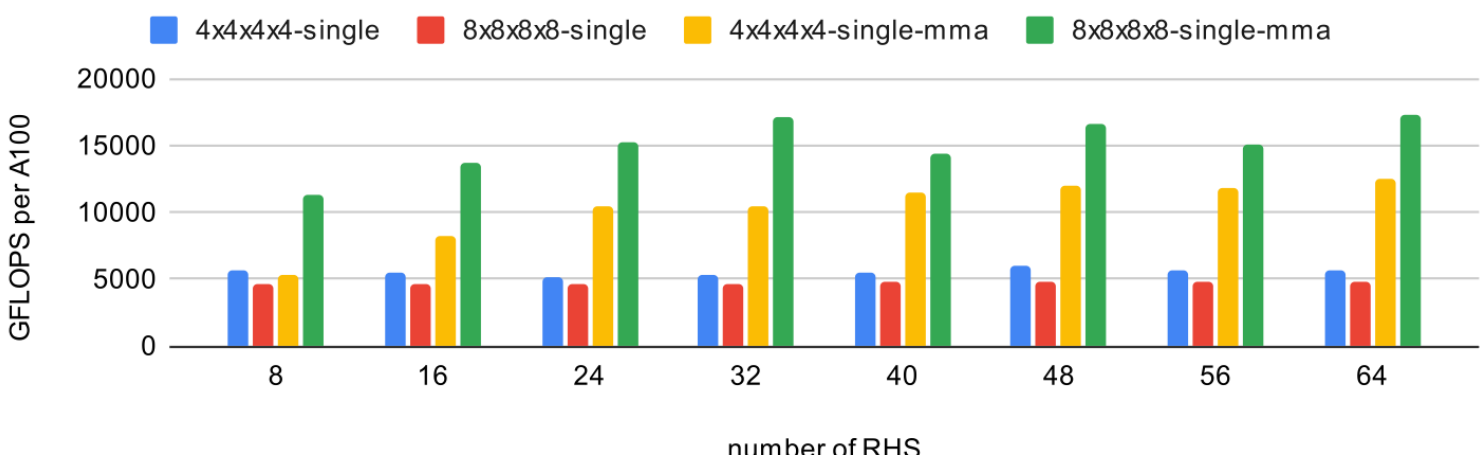
Results courtesy of Jiqun Tu

DL IS GOOD FOR HPC

Total performance of QUDA MatPCDagMatPC operator for 64^4 x 12 Mobius fermion on 8xH100-80GB-HBM3.
Yes, more than 100 TFLOPS in a DGX box.

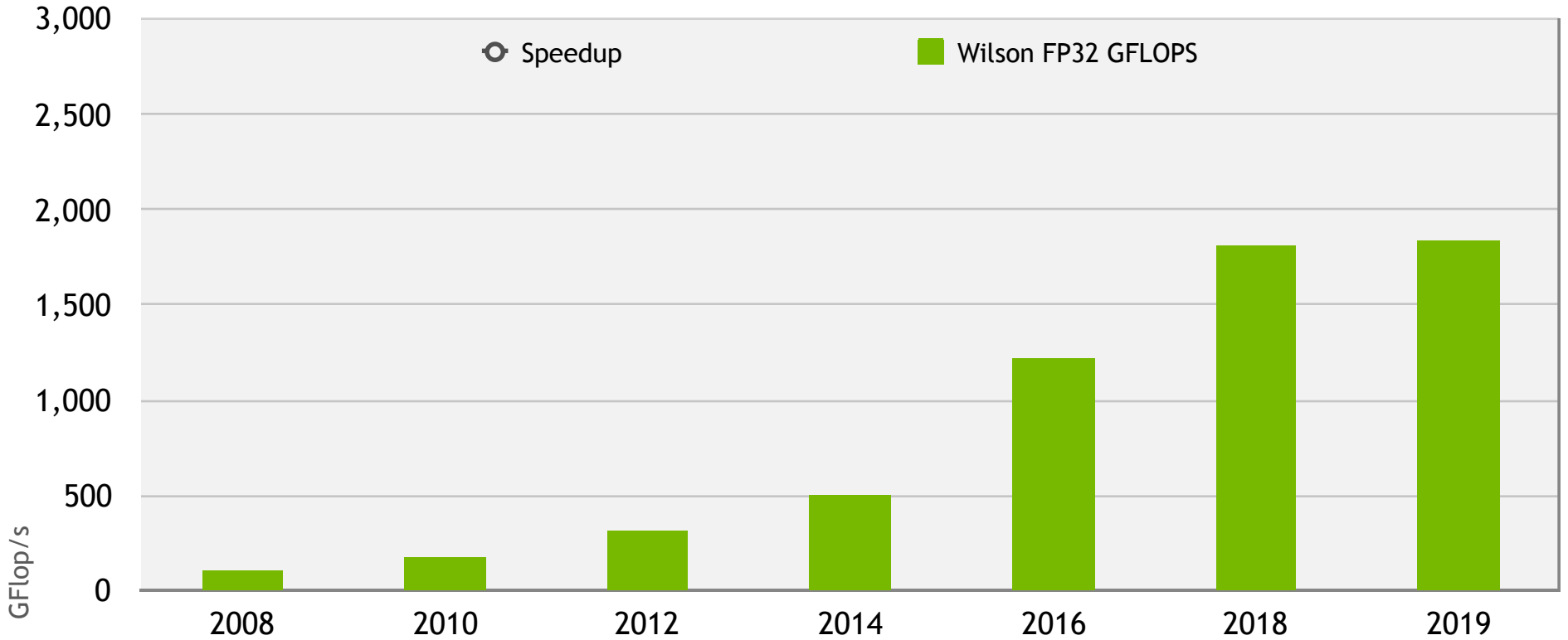
	QUDA performance [TFLOPS]	Equivalent Grid performance [TFLOPS]
Double	51.4	32.4
Single	105.8	66.7
Half (16-bit fixed-point storage with FP32 compute)	137.1	86.4

Coarse dslash with Nc = 24 utilize the tensor cores (MMA) on A100 GPUs in QUDA. See https://github.com/lattice/quda/pull/1355.



QUDA NODE PERFORMANCE OVER TIME

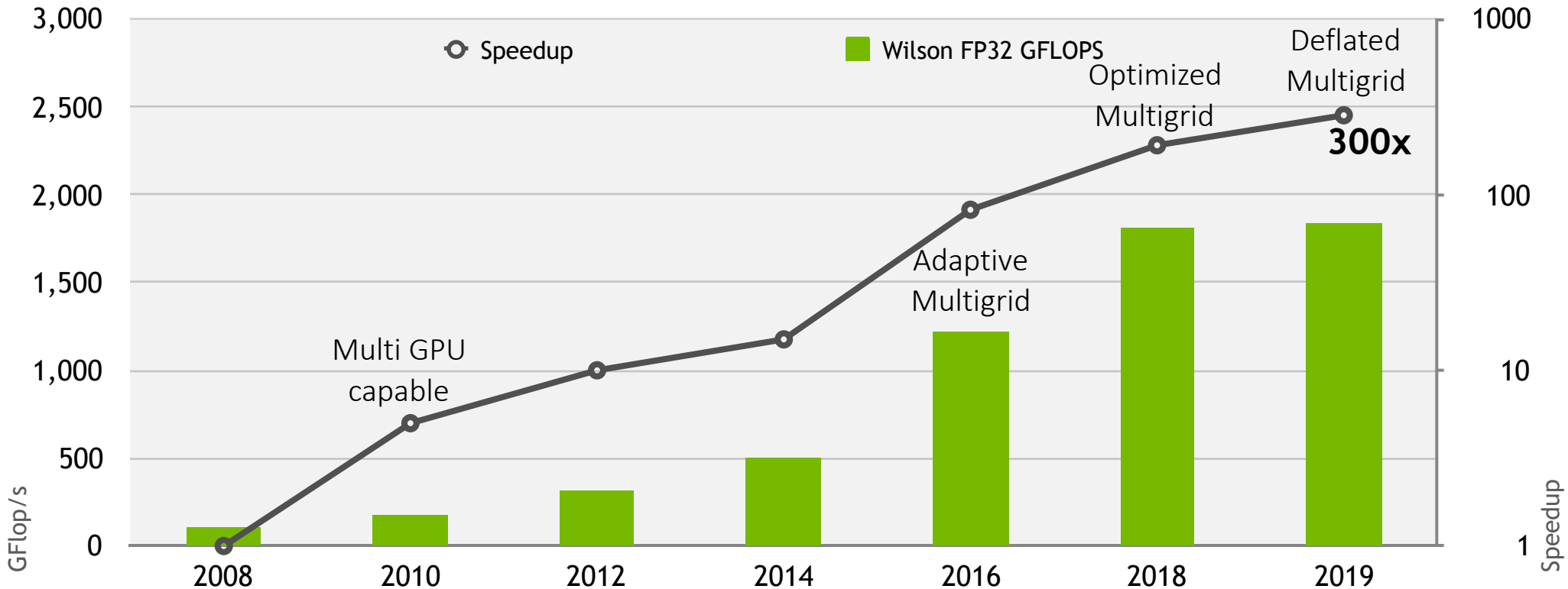
Multiplicative speedup through software and hardware



Speedup determined by measured time to solution for solving the Wilson operator against a random source on a $V=24^3 64$ lattice, $\beta=5.5$, $M_\pi = 416$ MeV. One node is defined to be 3 GPUs

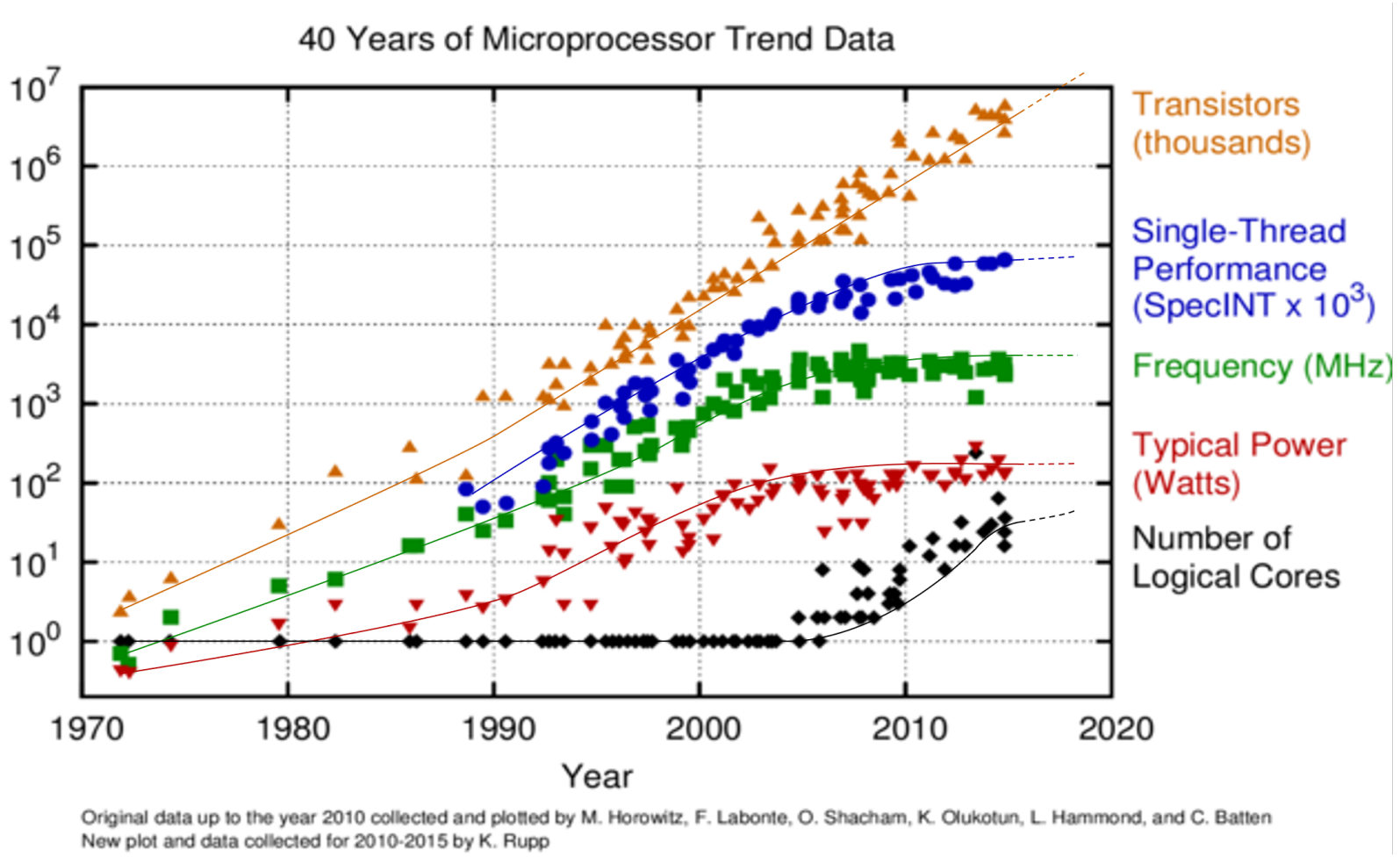
QUDA NODE PERFORMANCE OVER TIME

Multiplicative speedup through software and hardware



Speedup determined by measured time to solution for solving the Wilson operator against a random source on a $V=24^3 64$ lattice, $\beta=5.5$, $M_\pi = 416$ MeV. One node is defined to be 3 GPUs

PROTRACTED DEATH OF MOORE'S LAW

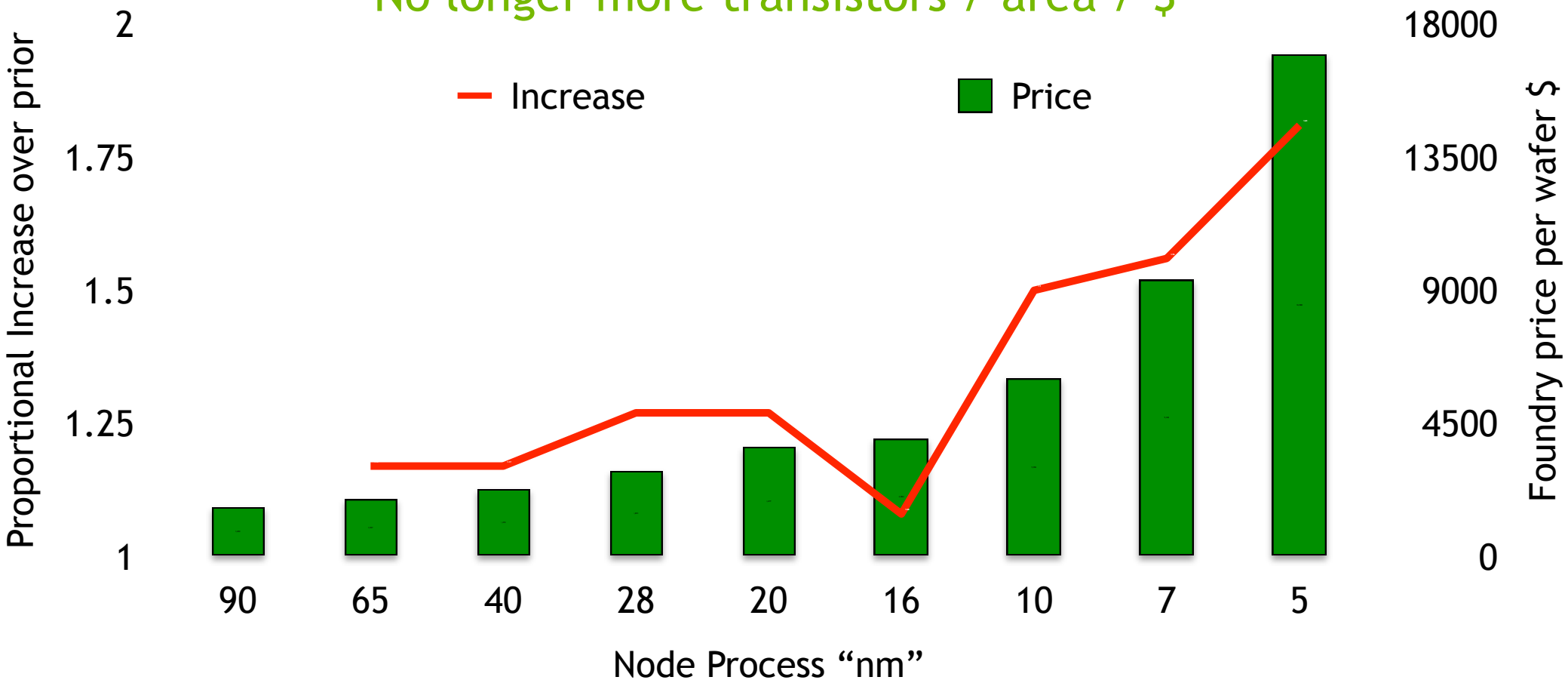


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

https://www.tomshardware.com/news/tsmcs-wafer-prices-revealed-300mm-wafer-at-5nm-is-nearly-dollar17000

MOORE'S LAW IS DEAD

No longer more transistors / area / $



NVIDIA.

DEATH OF MOORE'S LAW

CAP constrained

Cost

- Longer on a given process node
- Greater emphasis on microarchitecture innovation

Area

- Glue chips together "chiplets"
- Chip stacking

Power

- Liquid and immersion cooling
- Specialized instructions increasingly important (e.g., tensor cores)