Analysis on Energy Consumption in Computing aspect of Microelectronics

Sadasivan (Sadas) Shankar
Stanford SLAC and Materials Science & Engineering
Stanford, California

Semiconductor Industry Energy Efficiency Scaling (EES2) Technical Workshop

Hosted by the DOE/EERE Advanced Manufacturing Office (AMO)

September 14, 2022

Premise

- Environment
- Trends
  - Energy per Bit (EPB)
  - Energy per Instruction (EPI)
  - Energy per Application (EPA)
- Problem Trajectory
- Next Steps

Environment

Data Centers Are Facing a Climate Crisis

Companies are racing to cool down their servers as energy prices and temperatures soar. And the worst is yet to come.

WIRED

August 1, 2022

The World Meteorological Organization (WMO) says there's a 93 percent chance that one year between now and 2026 will be the hottest on record. Nor will that be a one-off. "For as long as we continue to emit greenhouse gases, temperatures will continue to rise,"

WHEN RECORD TEMPERATURES wracked the UK in late July, Google Cloud's data centers in London went offline for a day, due to cooling failures. The impact wasn't limited to those near the center: That particular location services customers in the US and Pacific region, with outages limiting their access to key Google services for hours. Oracle's cloud-based data center in the capital was also struck down by the heat, causing outages for US customers. Oracle blamed "unseasonal temperatures" for the blackout.

The UK Met Office, which monitors the weather, suggests that the record heat was an augur of things to come, which means data centers need to prepare for a new normal.

npr KQED

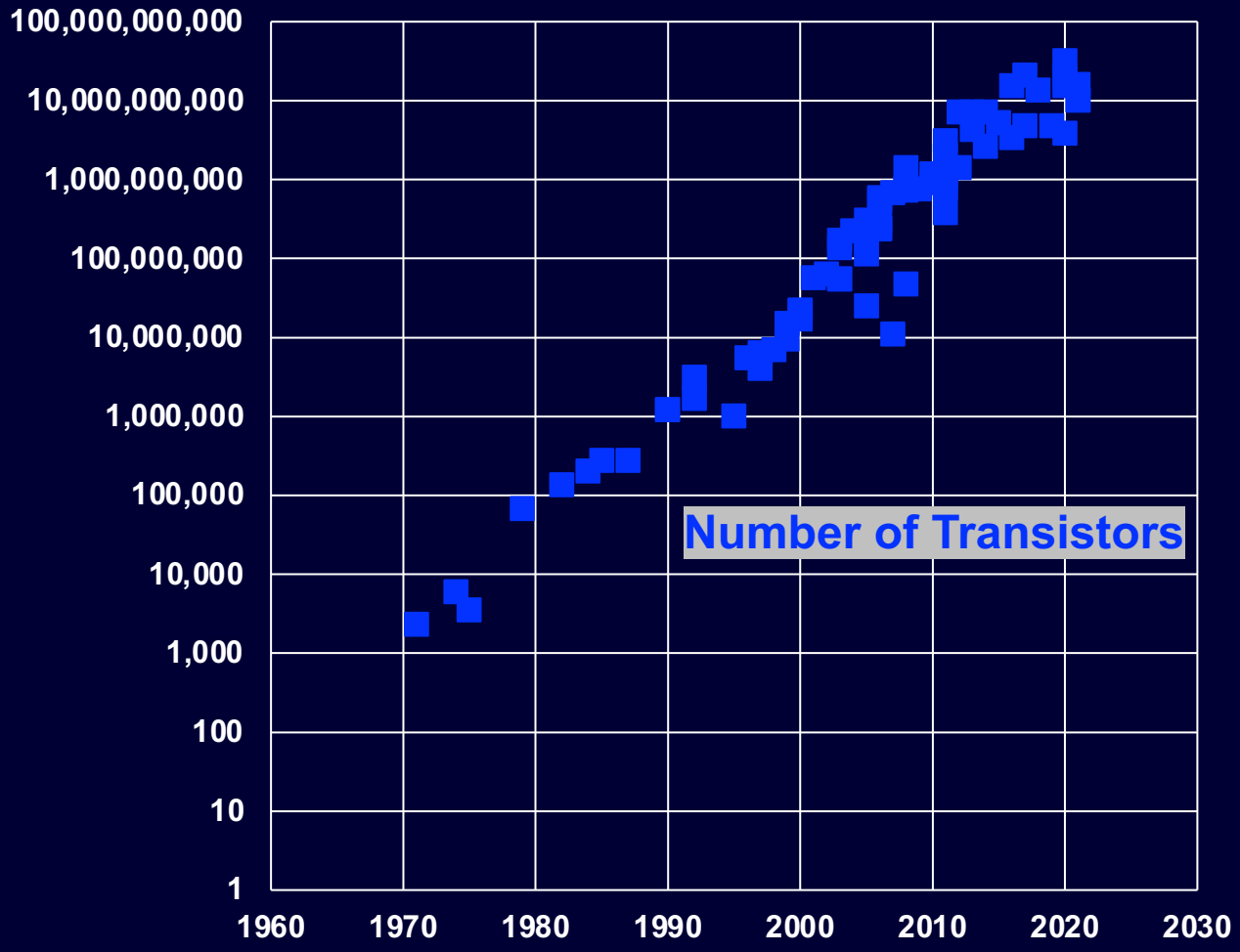
CLIMATE

The Arctic is heating up nearly four times faster than the whole planet, study finds

Temperatures in Longyearbyen, Norway above the Arctic Circle hit a new record above 70 degrees Fahrenheit in July 2020.
The Arctic has warmed nearly four times faster than the planet as a whole since 1979, a new study finds.

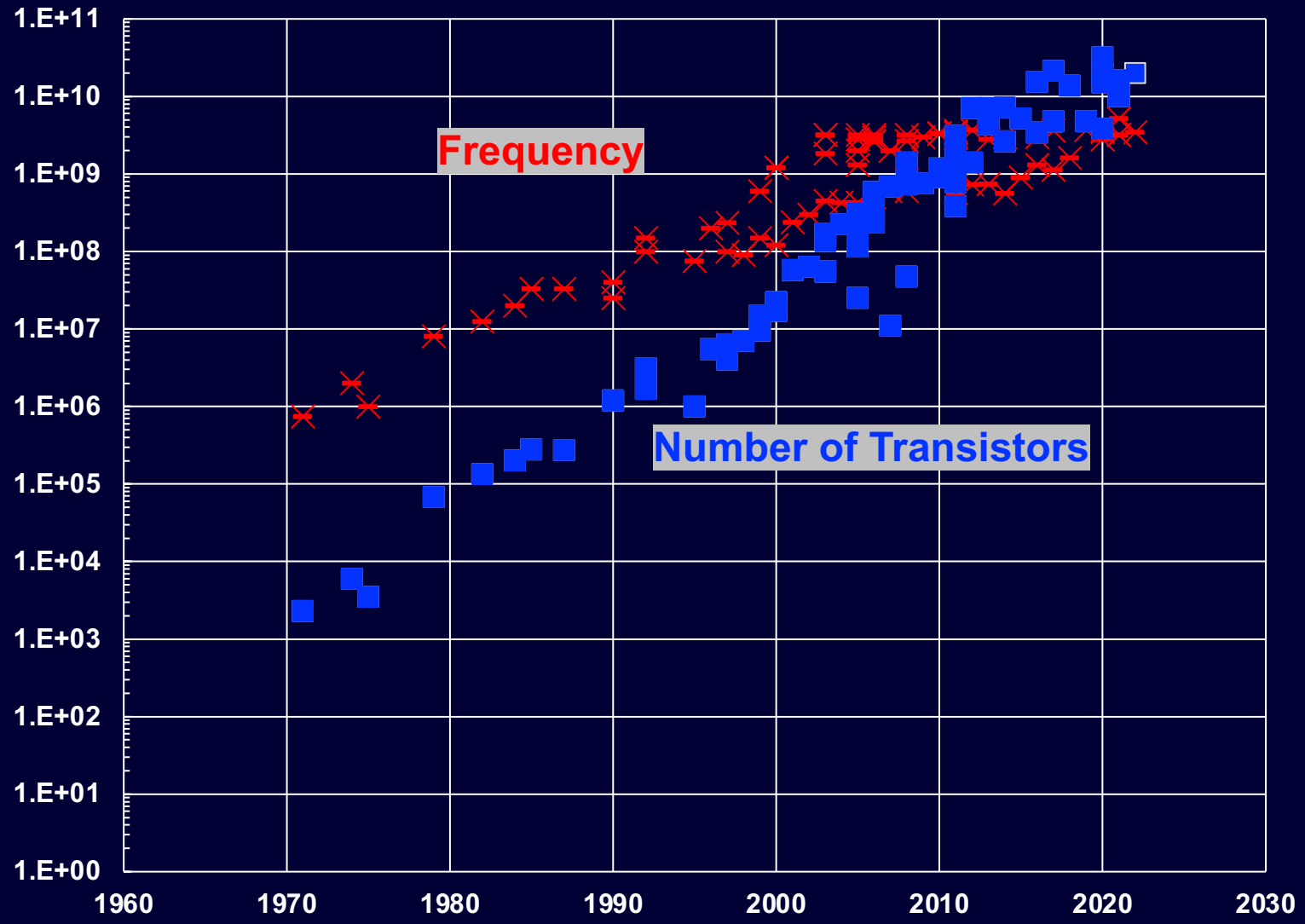
Energy per Bit (EPB)

Microprocessors (1970-2020)



- Geometric scaling over fifty years

Microprocessors (1970-2020)



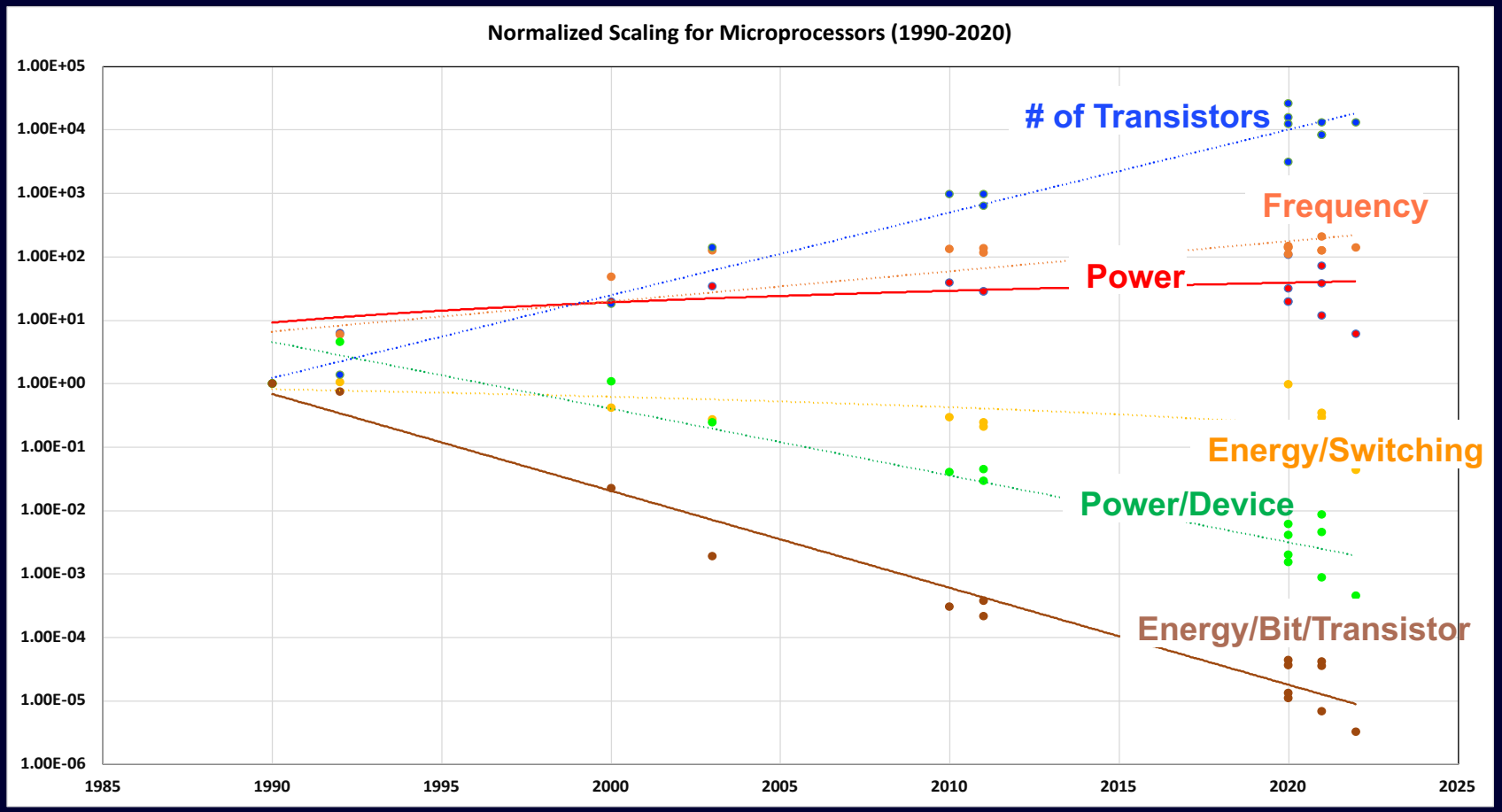
- Frequency flatter since 2000s

Microprocessors (1990-2020)

Product	Year of Introduction	Power (W)	Frequency (Hz)	Energy/Switching (Joules/switching or Joules/bit)*devices	Energy/Switching/ Transistor (Joules/switching/ Device OR Joules/bit)
Motorola 68040	1990	3.3	2.50E+07	1.32E-07	1.10E-13
DEC Alpha 21064 EV4	1992	21	1.50E+08	1.40E-07	8.33E-14
AMD Athlon	2000	65.7	1.20E+09	5.48E-08	2.49E-15
Pentium 4 Extreme Edition	2003	115	3.20E+09	3.59E-08	2.13E-16
Intel Core i7 Extreme Edition 980X (Hex core)	2010	130	3.33E+09	3.90E-08	3.34E-17
Intel Core i7 2600K	2011	95	3.40E+09	2.79E-08	2.41E-17
Intel Core i7 875K	2011	95	2.93E+09	3.24E-08	4.19E-17
AMD Ryzen 9 5900X	2020	105	3.70E+09	2.84E-08	1.48E-18
AMD Ryzen 3 3100	2020	65	3.50E+09	1.86E-08	4.89E-18
AMD Ryzen 3 3100 (fan outs)	2020	65	3.50E+09	1.86E-08	1.22E-18
Anton 3	2020	360	2.80E+09	1.29E-07	4.04E-18
Intel i9-12900K	2021	125	3.20E+09	3.91E-08	3.91E-18
Intel i9-12900K (Turbo)	2021	241	5.20E+09	4.63E-08	4.63E-18
Apple M1 Pro (10-core, 64-bit)	2021	39	3.20E+09	1.22E-08	7.62E-19
Apple M2 Pro (10-core, 64-bit)	2022	20	3.48E+09	5.75E-09	3.59E-19

- Number of transistors roughly follow Moore's law
- Latest Processor Switching energy is 0.36 atto joules/switching/transistor
- Significant benefits are coming from size scaling, while energy per switching is off by a factor of 25 over 32 years

Microprocessor (1990-2020): (Normalized)



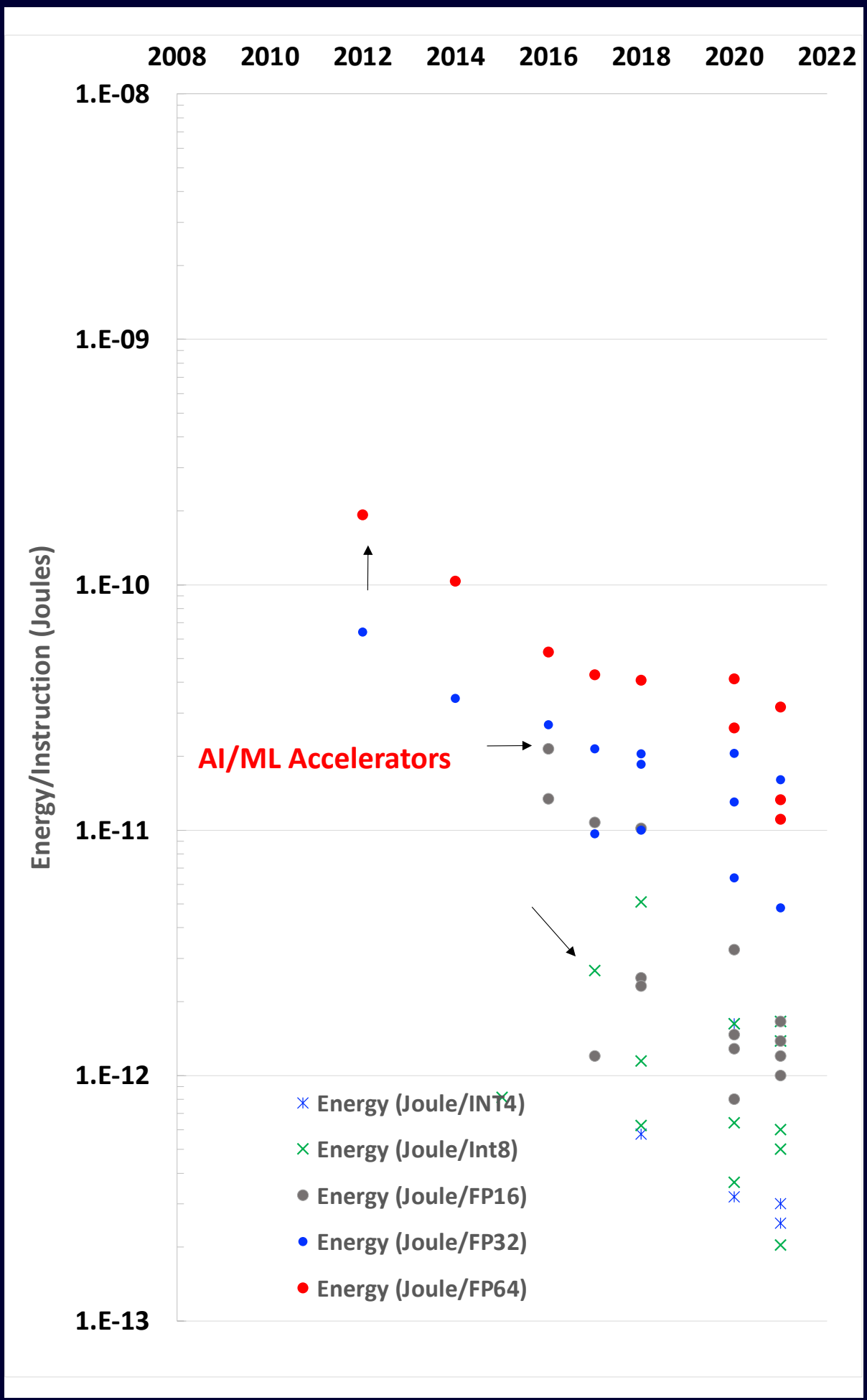
- Number of transistors roughly follow Moore's law
- Latest Processor Switching energy is 0.36 atto joules/switching/transistor
- Significant benefits are coming from size scaling, while energy per switching is off by a factor of 25 over 32 years

Energy per Instruction (EPI)

Basis of Analysis

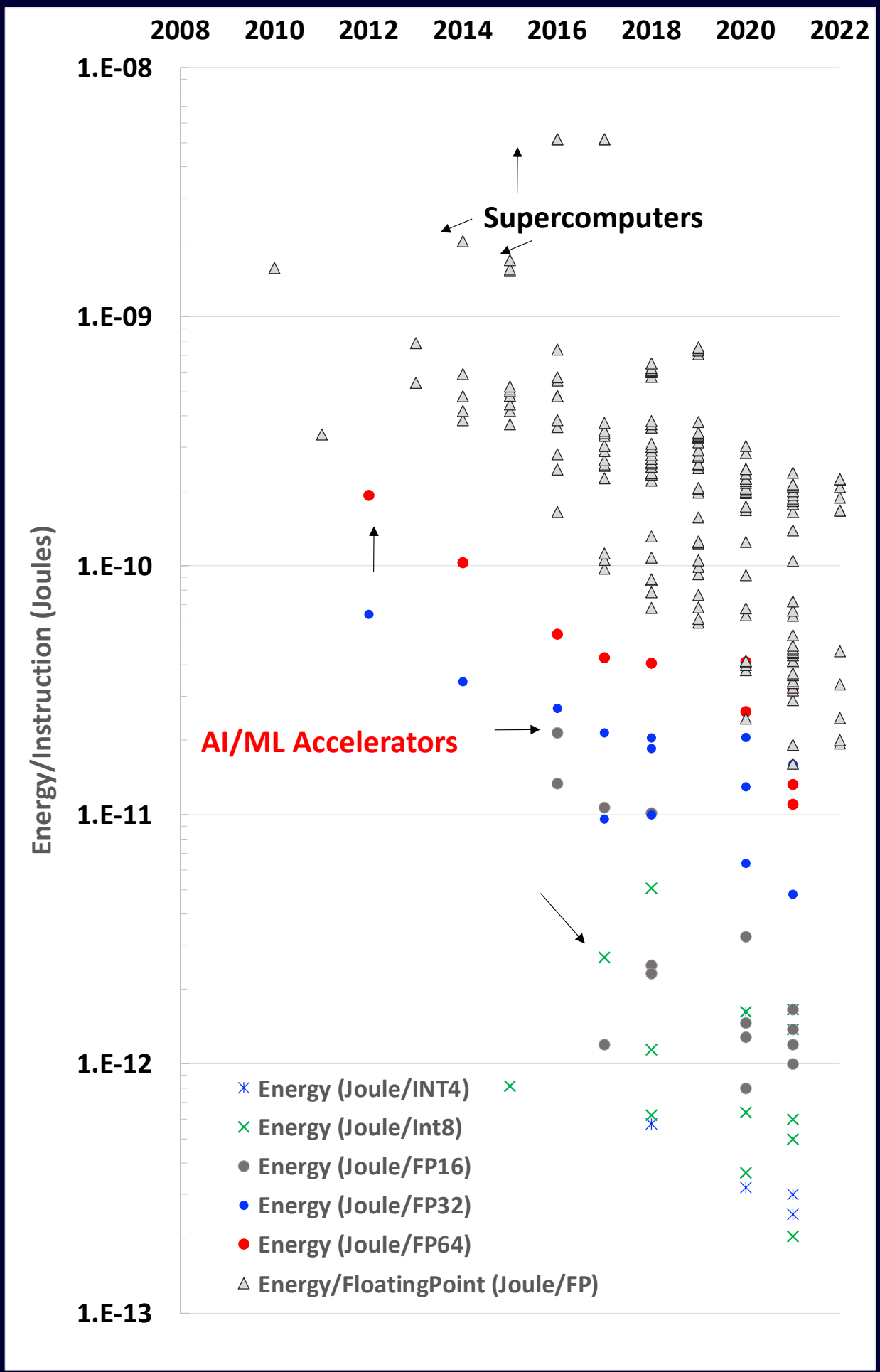
- **Two specialized computing systems analyzed: AI/ML Accelerators and Supercomputers**
  - Top down estimates of Energy per operation for different instructions (Int4, Int8, FP16, FP32, FP64)
  - Top500 Supercomputer list including the first exa-scale computer (HPL and HPCG)
- **Analysis are only estimates and help provide bounds and trends**
  - Data Analyzed from literature and shipped products and published work
  - Trends appear consistent across multiple sources

AI/ML Accelerators (2010-2020)



	Energy (Joule/Switching /Transistor)	Energy (Joule/INT4)	Energy (Joule/INT8)	Energy (Joule/FP16)	Energy (Joule/FP32)	Energy (Joule/FP64)
Minimum	2.1x10^-18	2.5x10^-13	2.0x10^-13	8.0x10^-13	4.8x10^-12	1.1x10^-11
Maximum	4.7x10^-17	1.6x10^-12	5.0x10^-12	2.1x10^-11	6.3x10^-11	1.9x10^-10
Maximum/Minimum	*22.75*	*6.64*	*24.95*	*26.71*	*13.30*	*17.42*

AI/ML Accelerators & Supercomputers

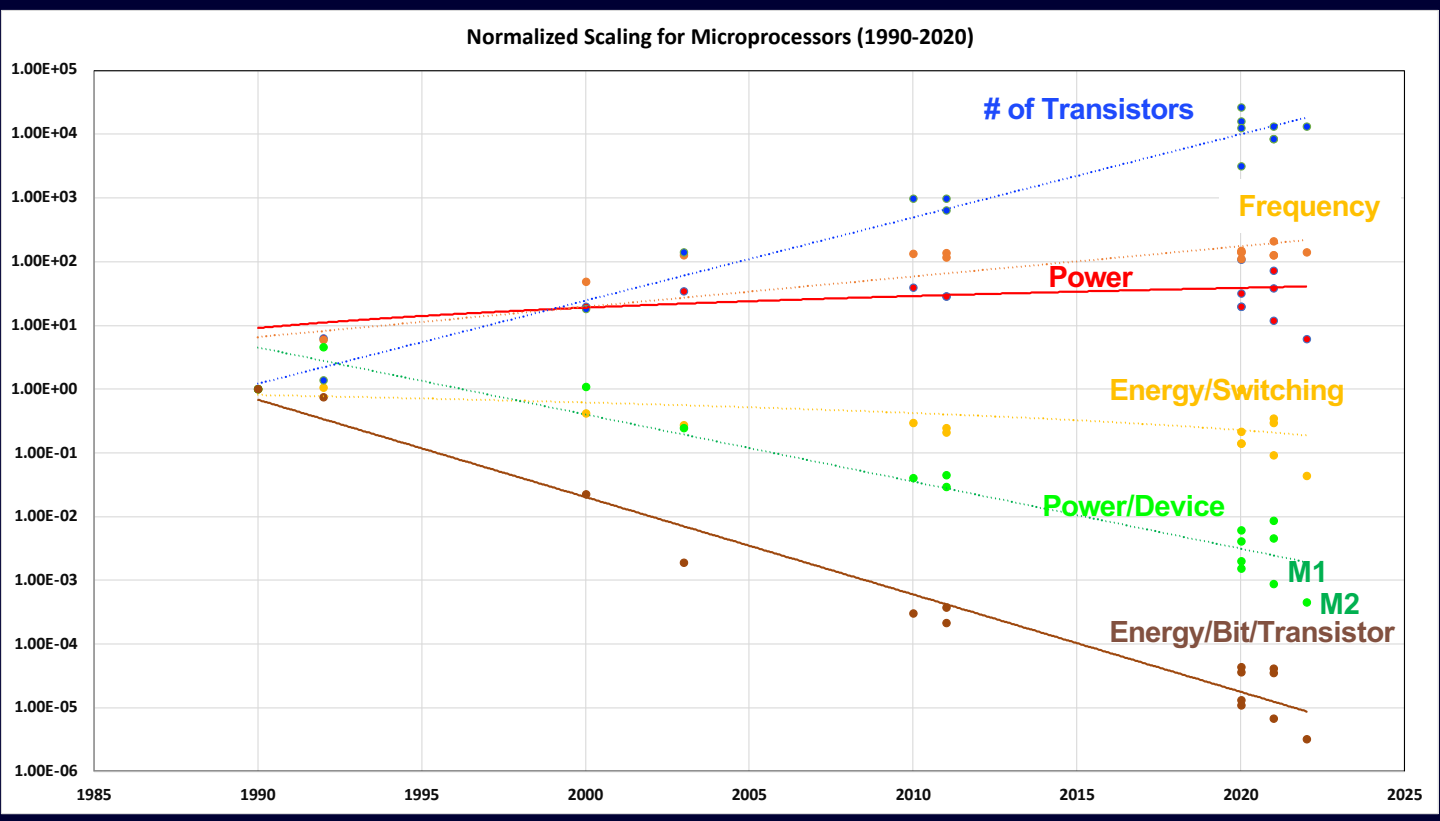


	Energy in SC (Joule/FP)
Minimum	1.6x10^-11
Maximum	5.1x10^-9
Maximum/Minimum	*321.81*

Energy per Application (EPA)

Machine Learning for NLP

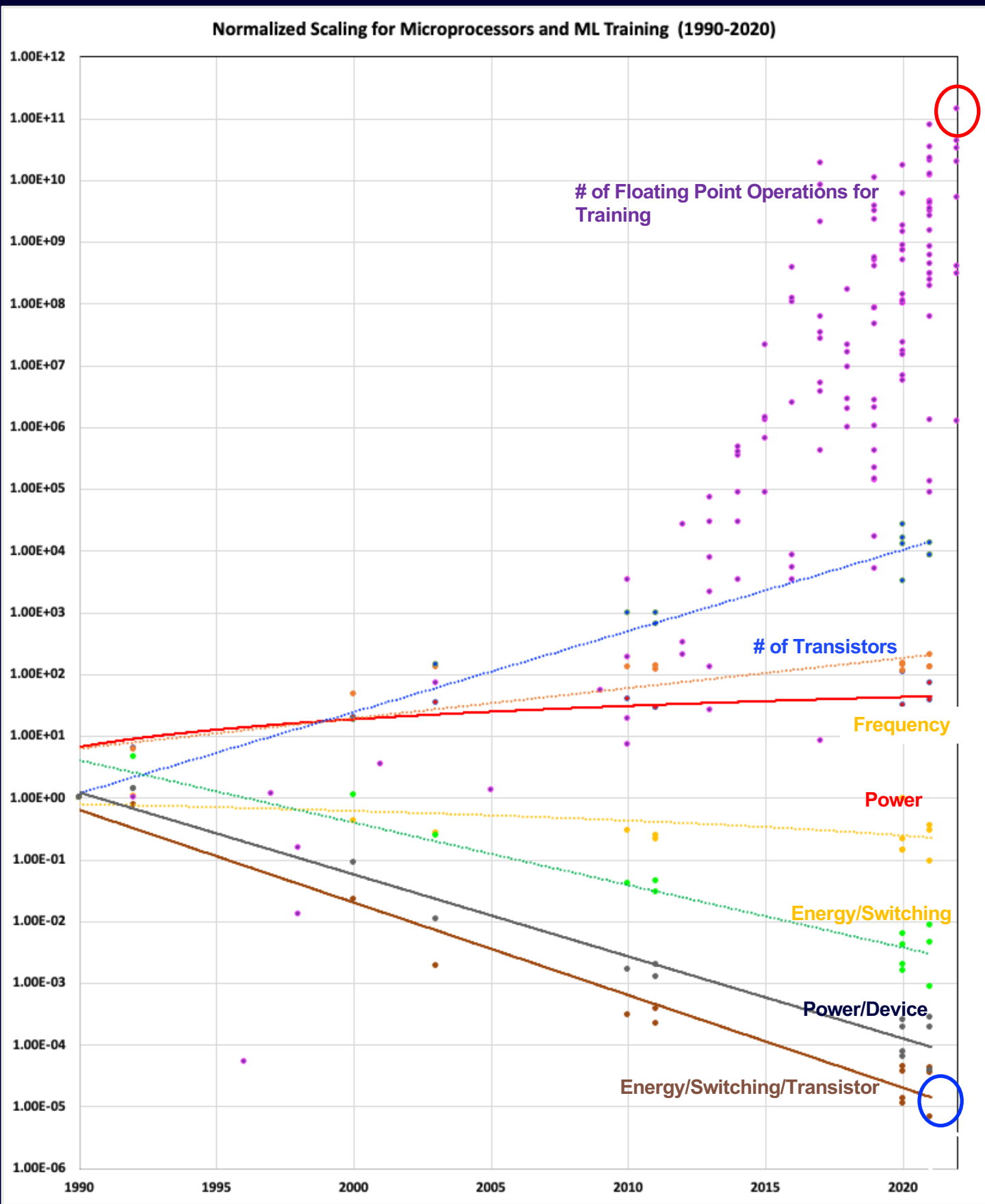
Normalized Scaling for Microprocessors (1990-2020)



- Normalized trends indicate that key attributes are largely driven by geometric scaling
- Recently energy efficiency scaling is being followed by one vendor (M1 and M2)

Normalized Scaling for Microprocessors (1990-2020)

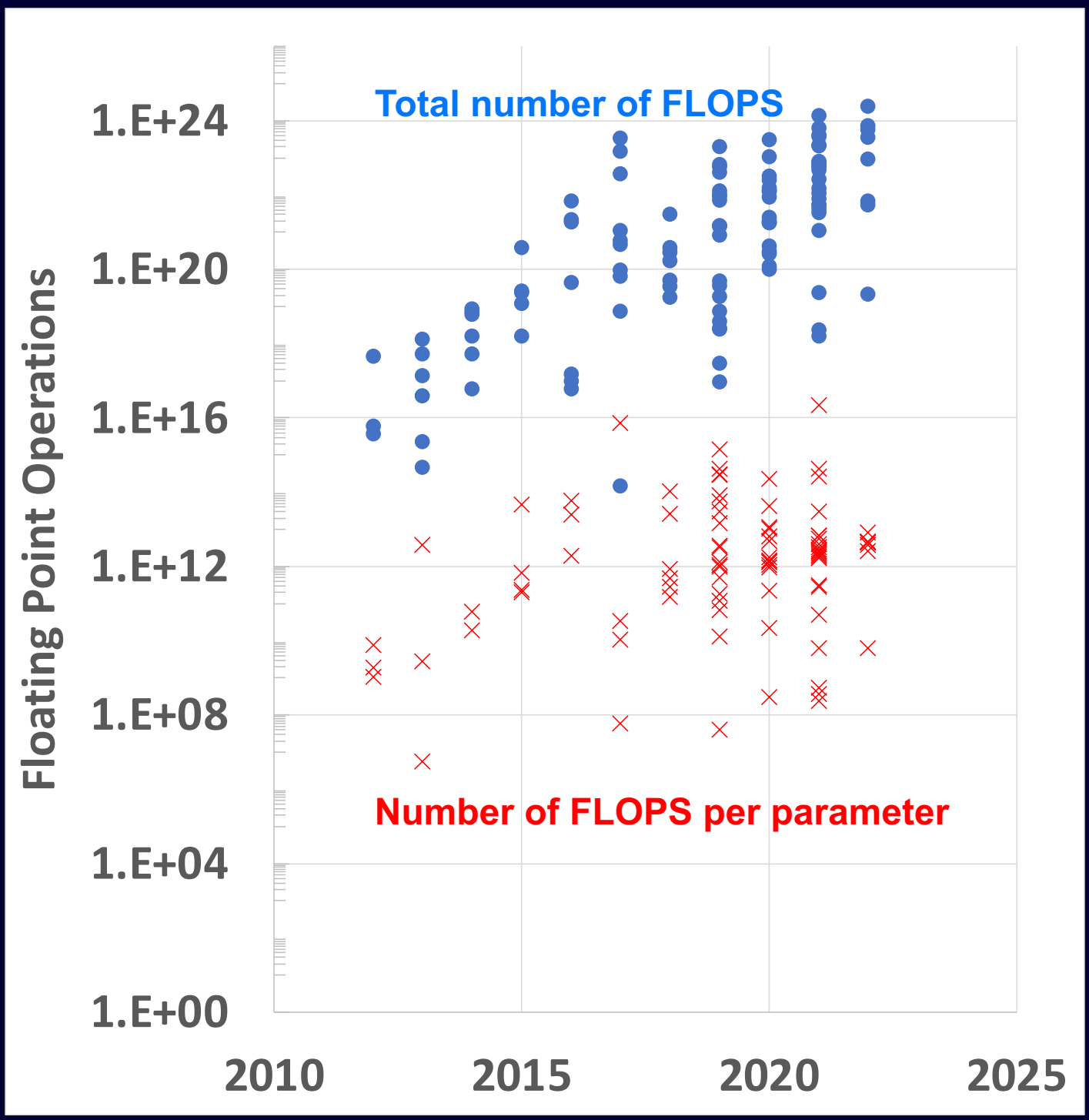
Sevilla, Jaime, et al. *arXiv:2202.05924* (2022).
Shankar, Reuther, submitted to IEEE'HPEC (2022)



AI/ML Training (1)

- Analysis indicates that the number of floating point operations for training Machine Learning are rapidly increasing
  - ~24 orders of magnitude over 70 years
  - ~16 orders of magnitude over 40 years
  - ~11 orders of magnitude over 30 years

AI/ML Training (2)



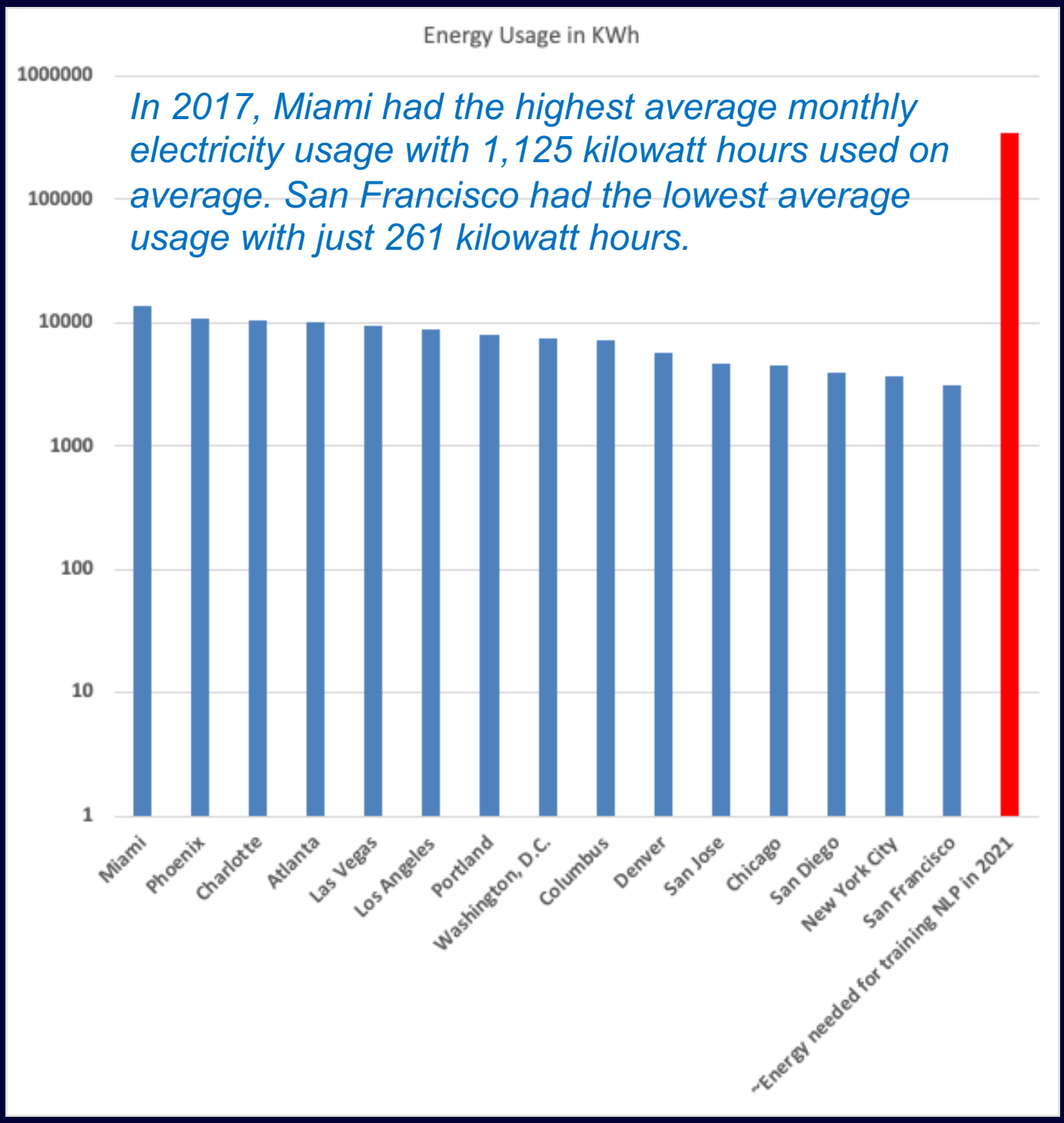
- Over 100 Million floating point operations per parameter

AI/ML Training (3)

- Analysis indicates that the number of floating point operations for training Machine Learning are rapidly increasing
  - ~24 orders of magnitude over 70 years
  - ~16 orders of magnitude over 40 years
  - ~11 orders of magnitude over 30 years
- Floating point operations for training Machine Learning are increasing faster than any reduction
- Average rough estimates:
  - Energy/FP operation ~ 1.0 x 10^{-12} Joules
  - Number of operations for training NLP between 2.2 x 10^{19} and 2.5 x 10^{24} (Average ~1+24)
  - Energy required = ~1.0 x 10^{12} Joules for training. = ~3.47 x 10^{5} Kwhr for training

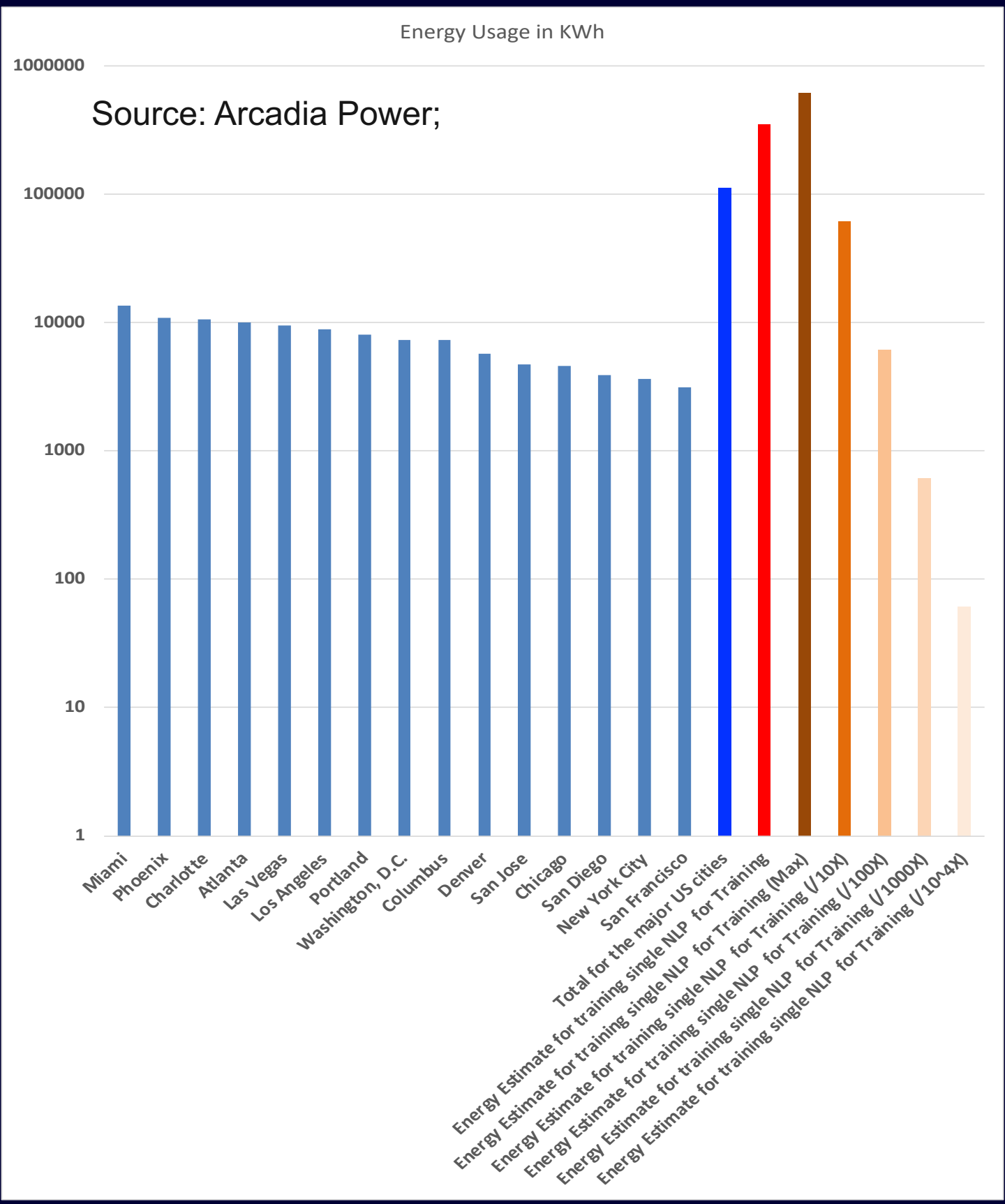
Annual Energy Usage in US Cities in 2017 compared to ML Training

Source: Arcadia Power
Shankar, Reuther, submitted to IEEE'2022



- Higher than the total monthly electricity usage of the 15 cities

Annual Energy Usage in US Cities in 2017 compared to ML Training



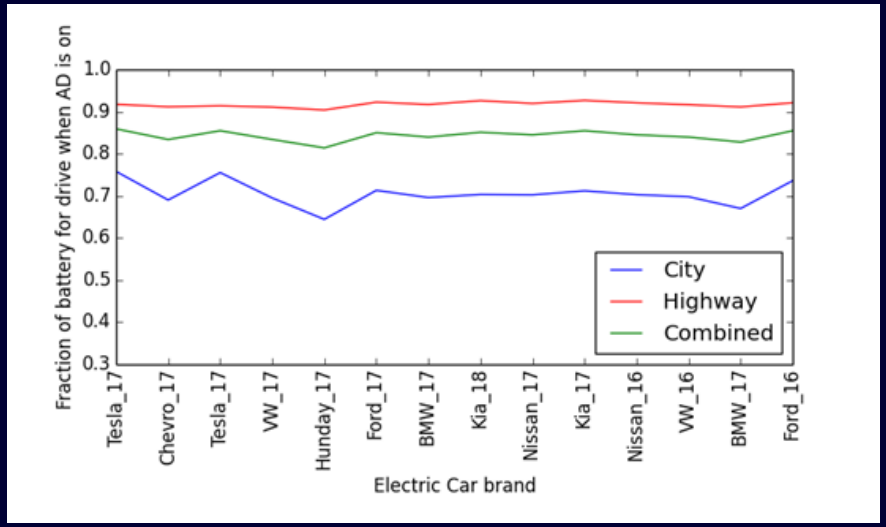
- 1000X reduction in ML usage brings the numbers to monthly electricity usage of the cities

Problem Trajectory

Energy Consumption: *Autonomous Cars*

Autonomous Cars' Big Problem:

Medium: May 15, 2019



The energy consumption of edge processing reduces a car's mileage with up to 30%.

S.Shankar

Energy Consumption: *Communication & Data Transfer*

TABLE I: Network and physical transport characteristics

Network Equipment					
Model	Router Type	B (Gb/s)	P_{max} (W)	E_{bit} (nJ/bit)	kJ/TB
Cisco CRS-3	Core	4480	12,300	2.7	21.6
Cisco 7609	Edge	560	4550	8.1	64.8
Cisco C3560CX 12PC-S (12 ports)	Layer 3 Switch	12×1	240	20	160

Physical Transport					
Type/Name	Weight (tons)	GVW (tons)	Load (tons)	Fuel E. (km/L)	kJ/kg/km
Bicyclist	0.075	-	0.05	-	2.56
Audi SUV	3.0	2.35	0.64	8.08	7.41
UPS Van	5.56	10.43	4.87	4.25	1.85
Delivery Truck	5.26	15.74	10.48	2.39	1.53
Boeing 757-200	-	116	39.8	0.17	5.66

Marincic, Foster, 2016

More efficient to transfer data in Packages than in Bits

Energy Consumption: *Cryptocurrency Mining*

Congress of the United States

Washington, DC 20515

July 15, 2022

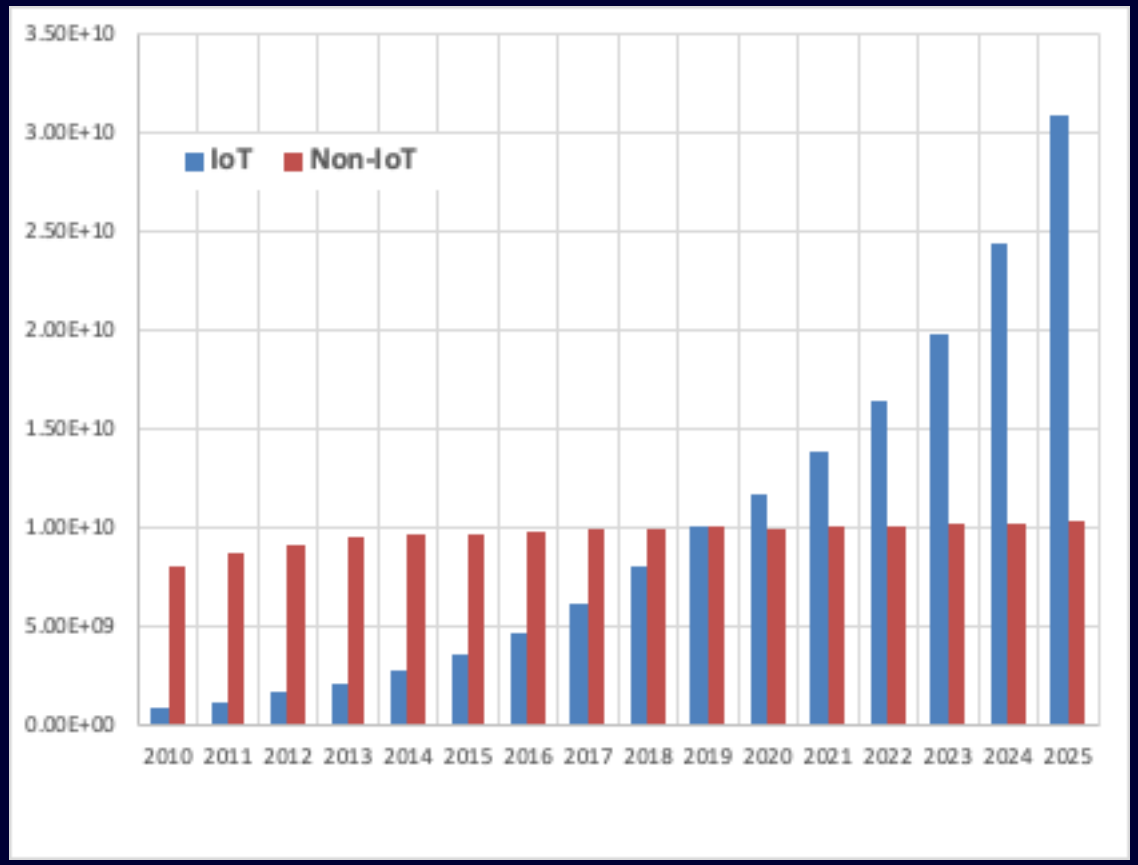
The cryptocurrency market has grown exponentially since first introduced over a decade ago.[1] Mining operations for Bitcoin, the largest cryptocurrency by market cap, are increasingly moving onshore, with the United States' share of global mining increasing from 4 percent in August 2019 to nearly 38 percent in January 2022 – meaning that over a third of the global computing power dedicated to mining Bitcoin is now drawn from miners in the U.S., in part due to a government crackdown in China last year.[2]

Cryptomining facilities' energy consumption is also causing significant increases in energy costs for many small businesses and residents. Cryptomining in the city of Plattsburgh, New York reportedly resulted in residential electricity bills that were "up to $300 higher than usual" in the winter of 2018, leading the city to introduce the nation's first 18-month moratorium on new cryptomining operations.[8] A recent study estimates that "the power demands of cryptocurrency mining operations in upstate New York push up annual electric bills by about $165 million for small businesses and $79 million for individuals."[9] Moreover, states like Texas with relatively cheap electricity costs are experiencing an influx of cryptomining companies, raising concerns about the state's unreliable electricity market and the potential for cryptomining to add to the stress on the state's power grid.[10]

Energy costs runover driven by computations for "mining"

Number of Devices

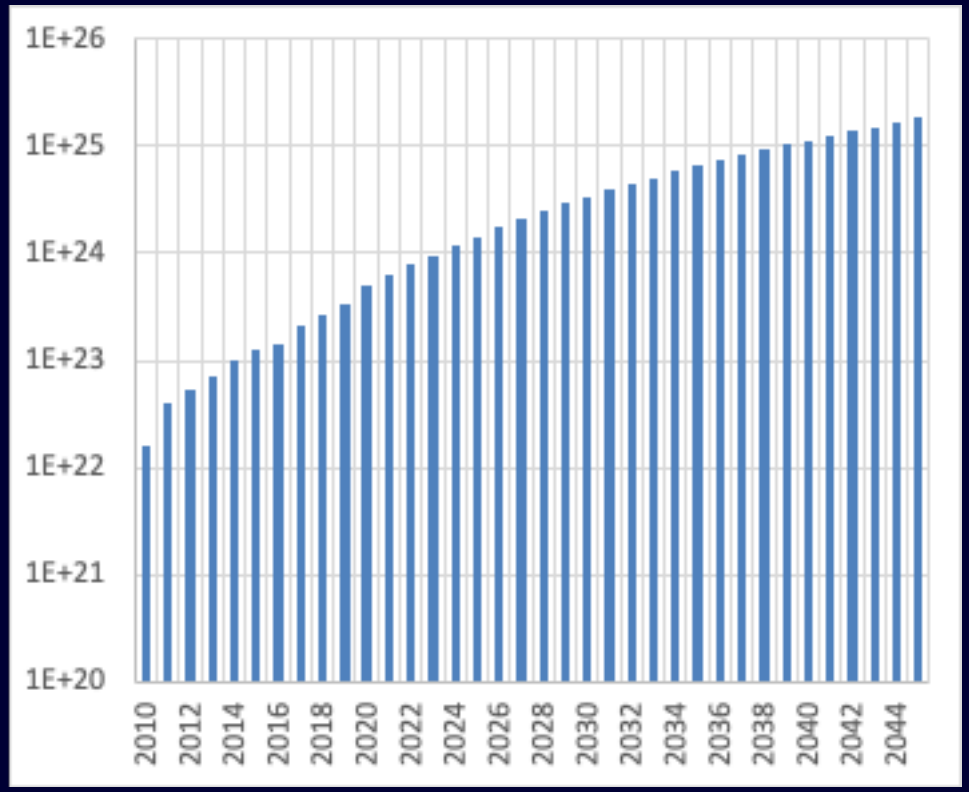
IoT Analytics, 2020



Total devices doubling by 2025 (~2.25% of World Power of 17.7 TW)

Amount of Data

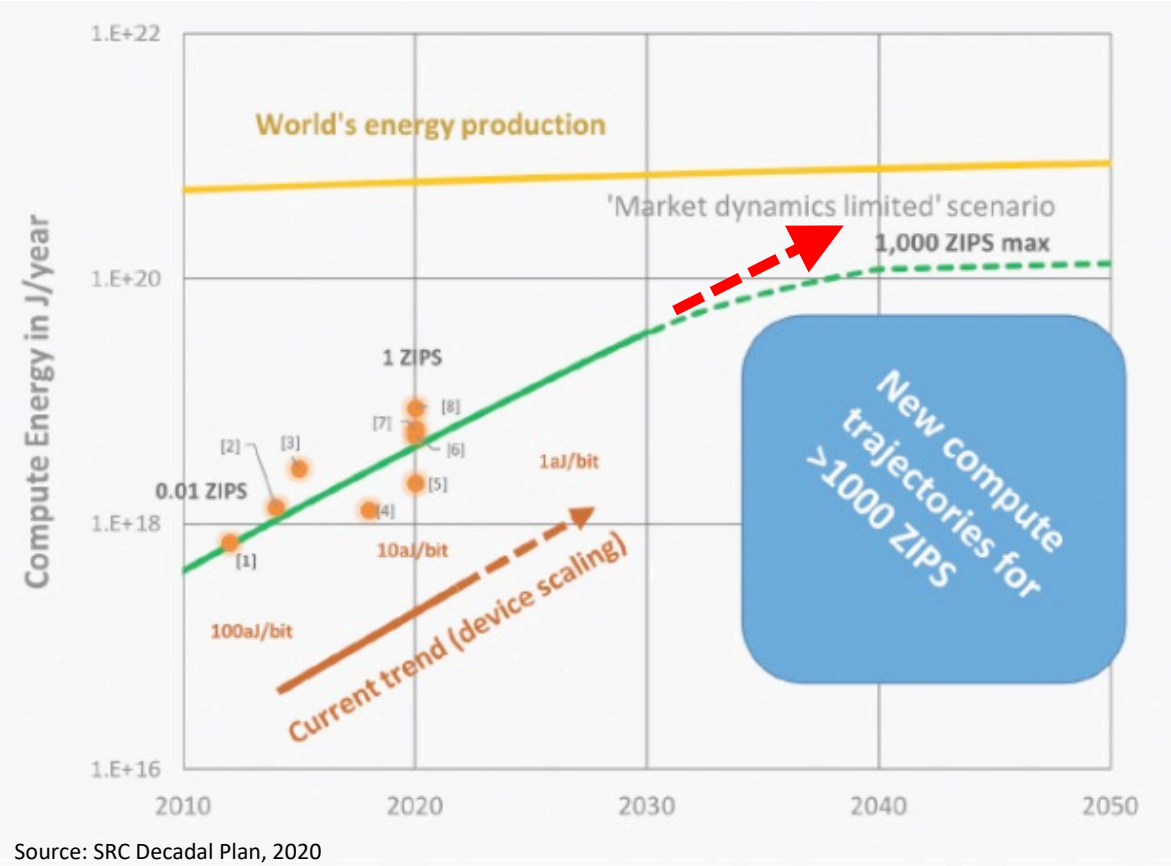
IDC, 2021



Computing and Data Storage (~2% of World Needs) both need energy

Unsustainable Computing Energy Trajectory

Seismic shift #5: Computing energy is not sustainable



Source: SRC Decadal Plan, 2020

Why Seismic Shift?

Computing will not be sustainable by 2040, as its energy requirements would exceed the estimated world's energy production

Need: Discover computing paradigms/architectures with a radically new 'computing trajectory' demonstrating >1,000,000x improvement in energy efficiency. Changing the trajectory not only provides immediate improvements but also provides many decades of buffer and is much more cost effective than attempting to increase the world's energy supply dramatically.

Next Steps

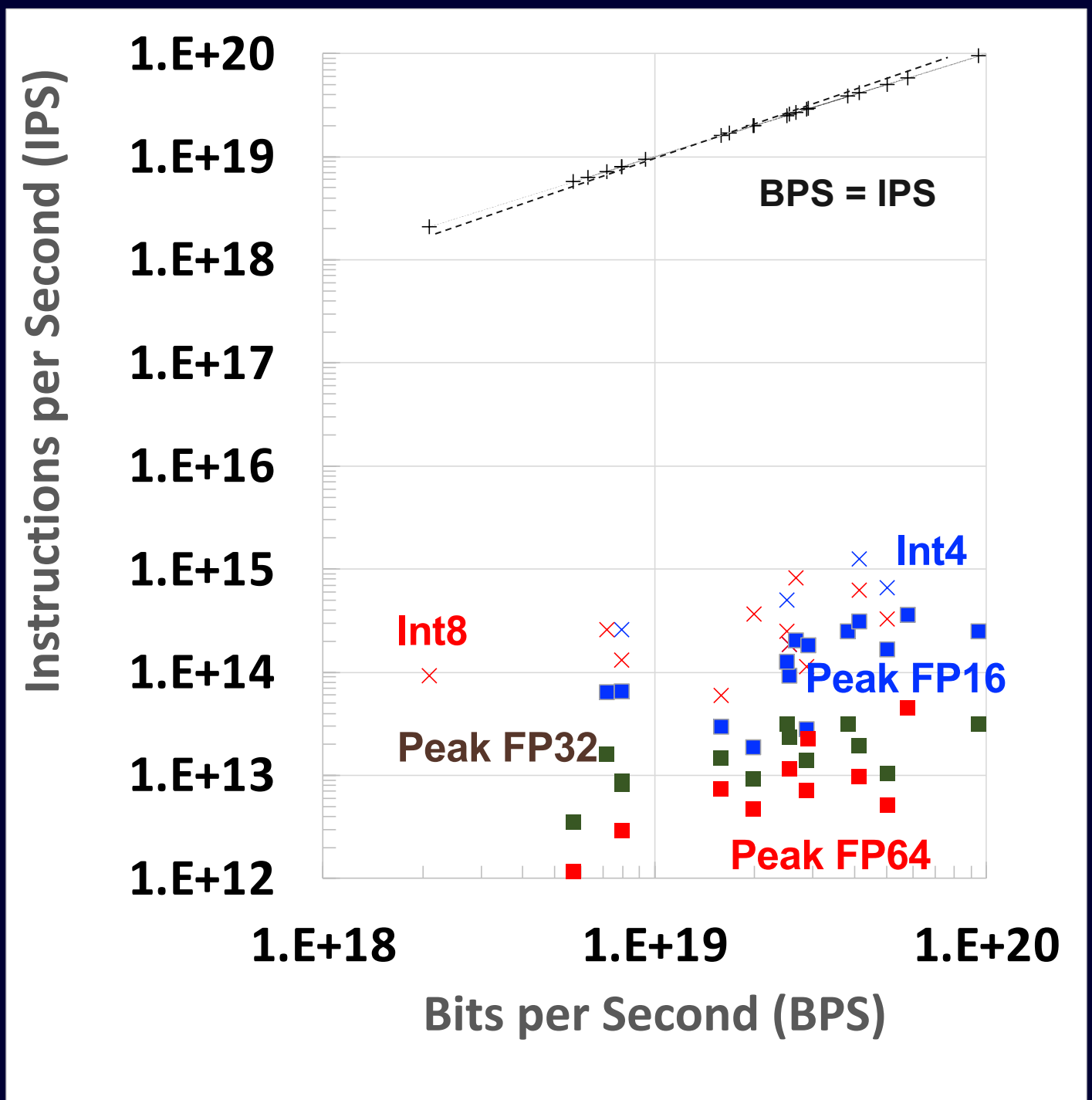
BIT Utilization (1)

- Bits and Instructions
  - The number of bits switching per second relates to the frequency related to the switching rate of all the transistors
  - At the system level, the corresponding variable is the *number of instructions per second*

- BIT Utilization
  - Instructions per second (IPS) for the system are the same as the number of bits switching for all the transistors (BPS), then the bit utilization (BPS/IPS) will be unity. => all bits are proportionally utilized for system level instructions

BIT Utilization

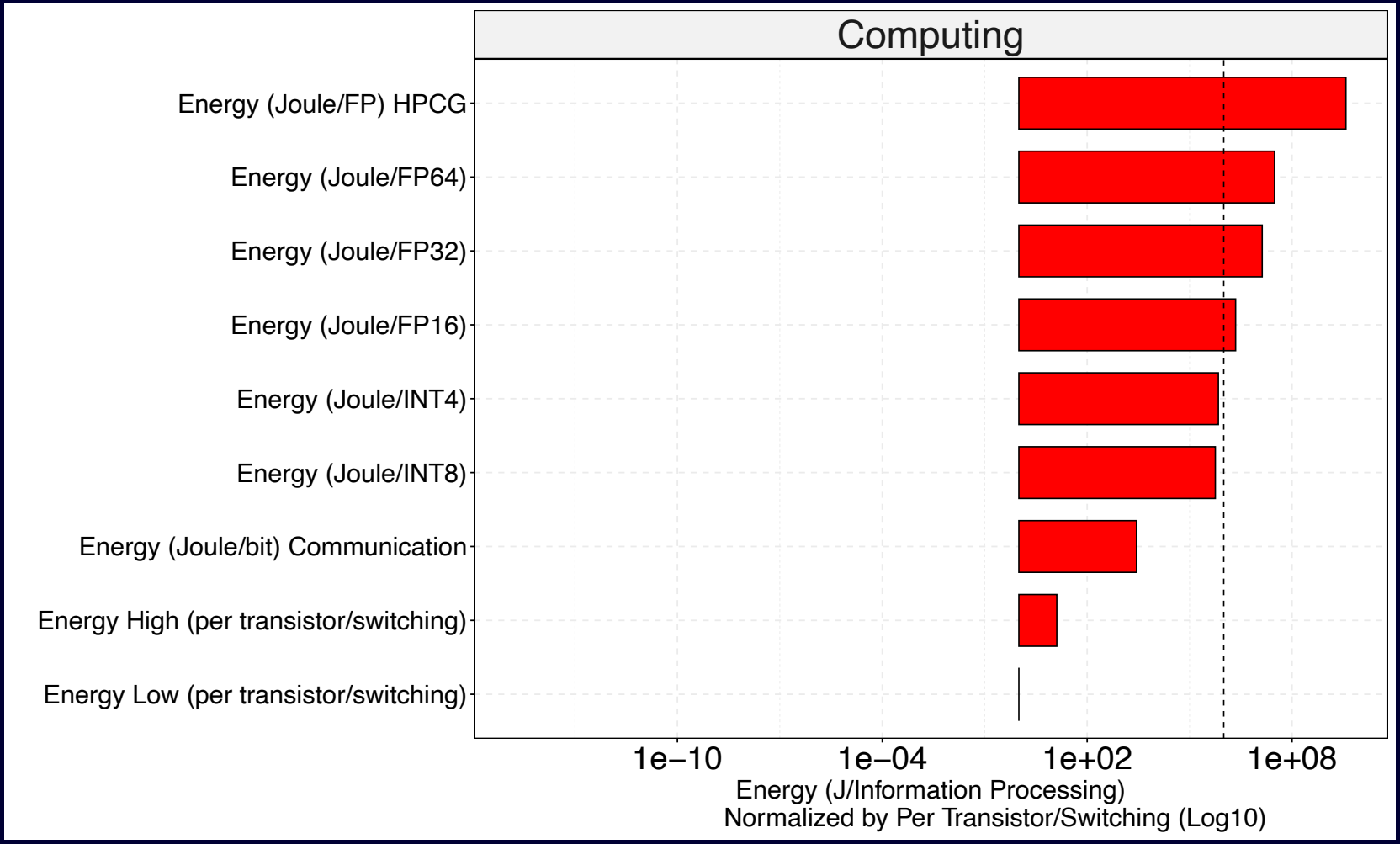
Shankar, Reuther, submitted to
IEEE'HPEC (2022)



- Int4 has higher bit utilization
- FP64 has lower bit utilization

Map for Energy Efficiency (0)

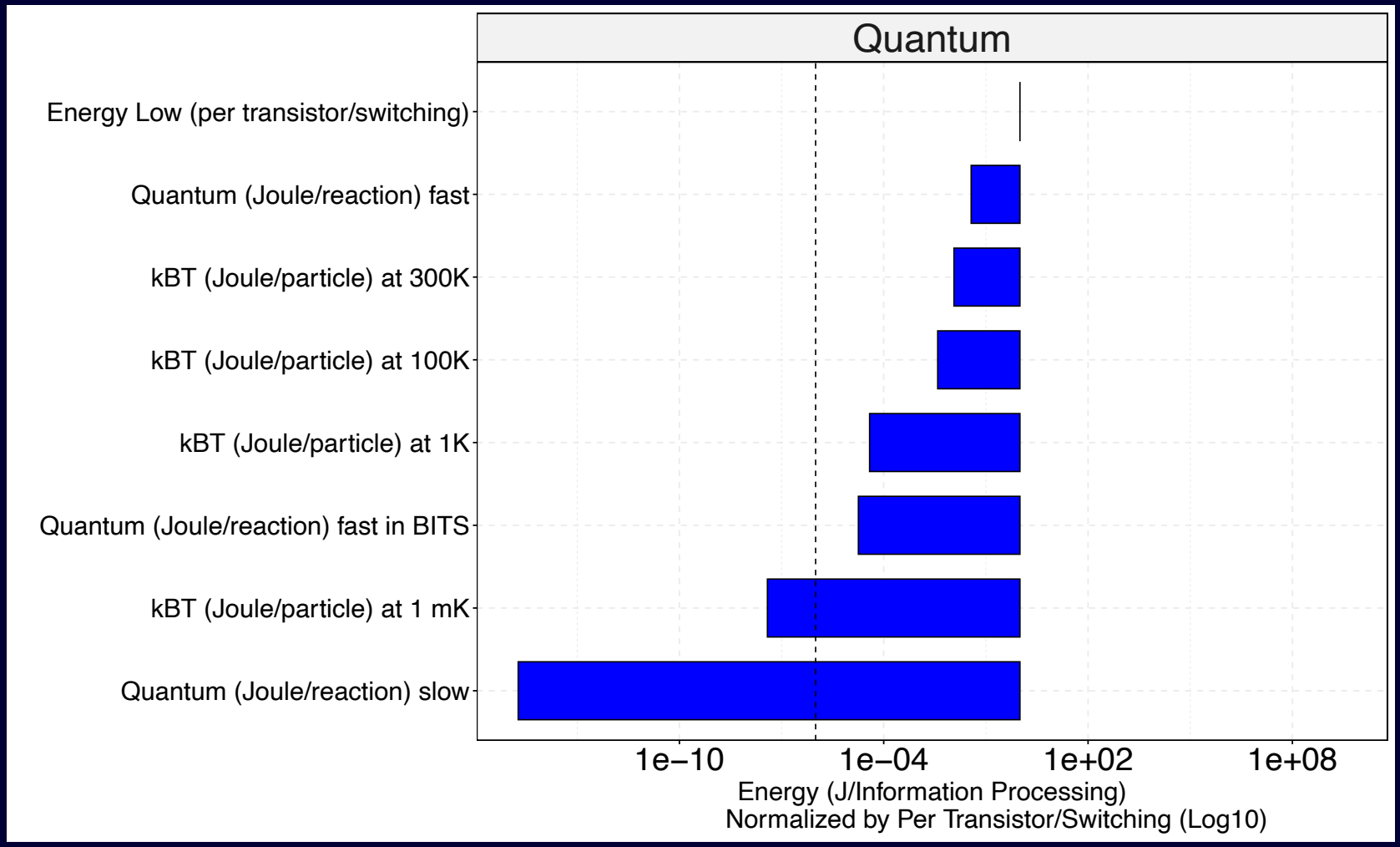
Map for Energy Efficiency (1)



Ack: V. Shankar

- Normalized with respect to Energy/transistor/switching
- Compared to baseline, System shows about 8 orders of magnitude higher (*does not include application-specific metrics*)

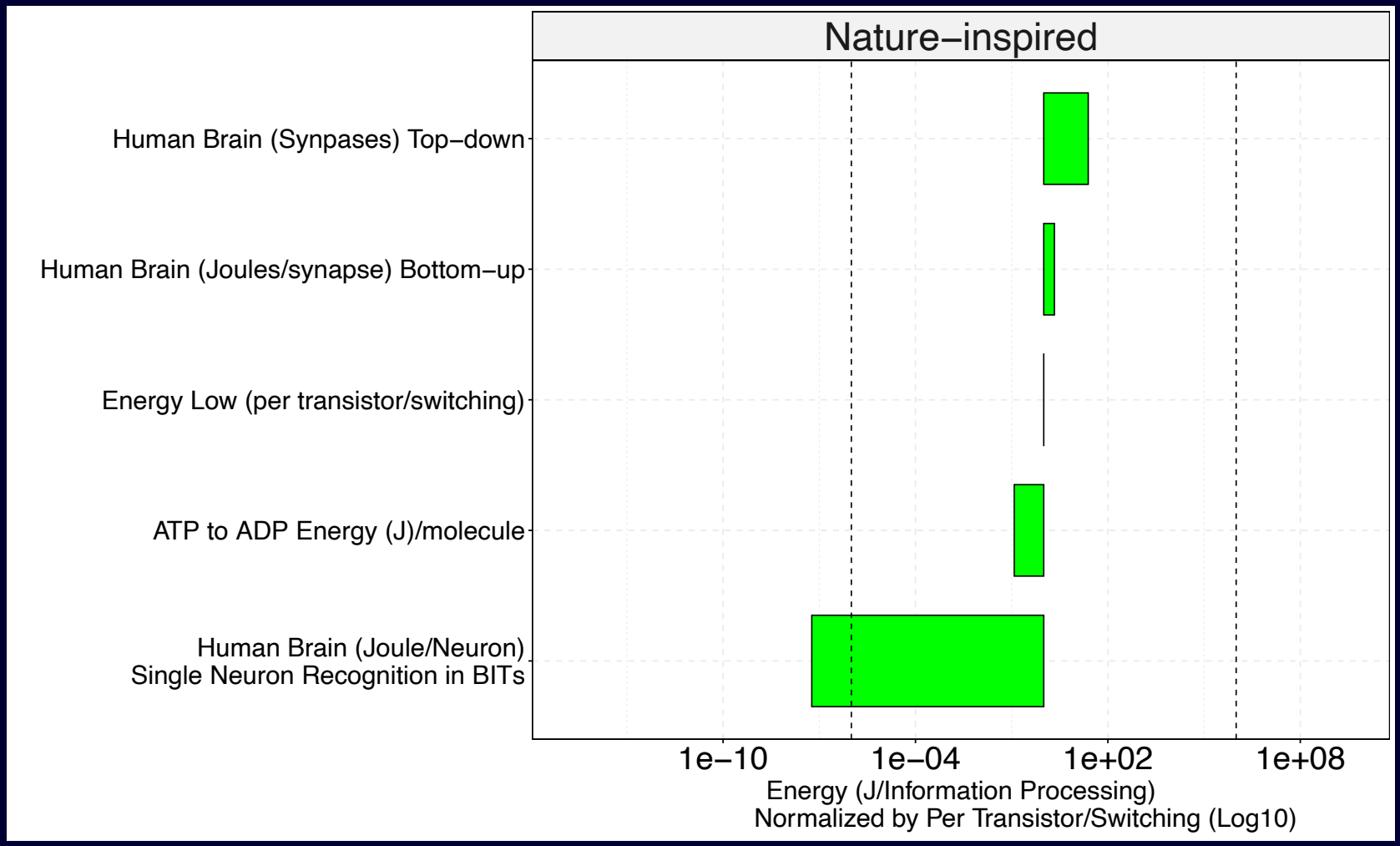
Map for Energy Efficiency (2)



Ack: V. Shankar

- Normalized with respect to Energy/transistor/switching
- Compared to baseline, Quantum-level fast chemical reaction shows about 4-7 orders of magnitude lower

Map for Energy Efficiency (3)

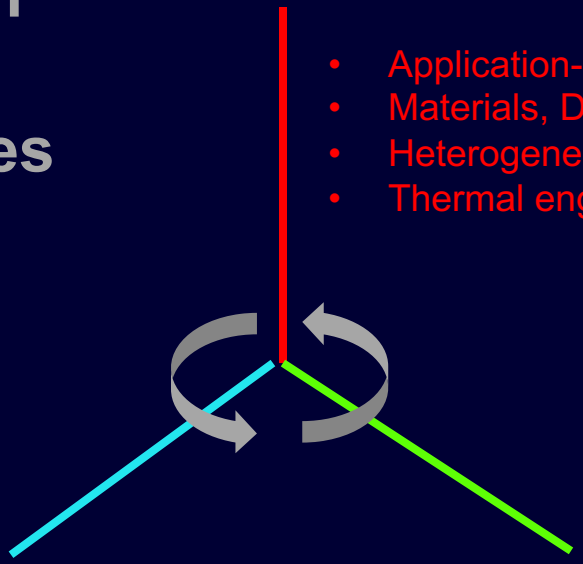


Ack: V. Shankar

- Normalized with respect to Energy/transistor/switching
- Compared to baseline, Single neuron/synapse switching shows about 6-8 orders of magnitude lower

Map for Energy Efficiency (4)

Mix and Match among the 3 different lines



System

- Application-specific Architectures, Algorithms
- Materials, Devices
- Heterogeneous Integration/Packaging
- Thermal engineering

Quantum

- Cryogenic engineering
- Error Correcting Devices, Hardware, Algorithms
- Application-specific Q-Information Processing
  - Fermionic Quantum Computing
  - Mixed states
- Efficient Q-C/C-Q Converters

Nature-inspired

- Bottom-up Processing
- Probabilistic and/or Statistical Computing
- Fractal Architectures
- Application-specific Informational basis, Hardware
  - “BIT is more than a bit”

Summary

- Energy Efficient Scaling in computing is necessary for both sustainability and ability to solve realistic problems
- Energy Efficiency Scaling consists of three overlapping components
  - Energy/Bit
    - (Materials, Devices)
  - Energy/Instruction
    - (Architecture, Integration, System, Devices)
  - Energy/Application
    - (Algorithms x Software)
- Multiple innovations at multiple-levels can enable EES
  - Headroom for EES exists; Many of the speakers will address
- Current challenges are opening new pathways to an exciting computing future!!!

Acknowledgements and References

Thanks!

- Partially funded by the U.S. Department of Energy's Office of Science contract DE-AC02-76SF00515 with SLAC (Ack: C.C. Kao, P. McIntyre)
- R. Zare (Stanford), W. Goddard (Caltech)
- A. Reuther (MIT-LL)
- V. Zhirnov (SRC), M. Khairy and T. Rogers (Purdue), T. Besiroglu (MIT)

Papers in Progress

1. Shankar, S. 2021 "*Lessons from Nature for Computing: Looking beyond Moore's Law with Special Purpose Computing and Co-design",* 2021 IEEE High Performance Extreme Computing Conference (HPEC) (pp. 1-8).
2. Shankar, S, Reuther, A, 2022, "*Trends in Energy Estimates for Computing in AI/Machine Learning Accelerators, Supercomputers, and Compute-Intensive Applications*", 2022 IEEE High Performance Extreme Computing Conference (HPEC)
3. *A Logical Framework for Information Processing* (in preparation)
4. *Energy-based Scaling for Computing* (in preparation)