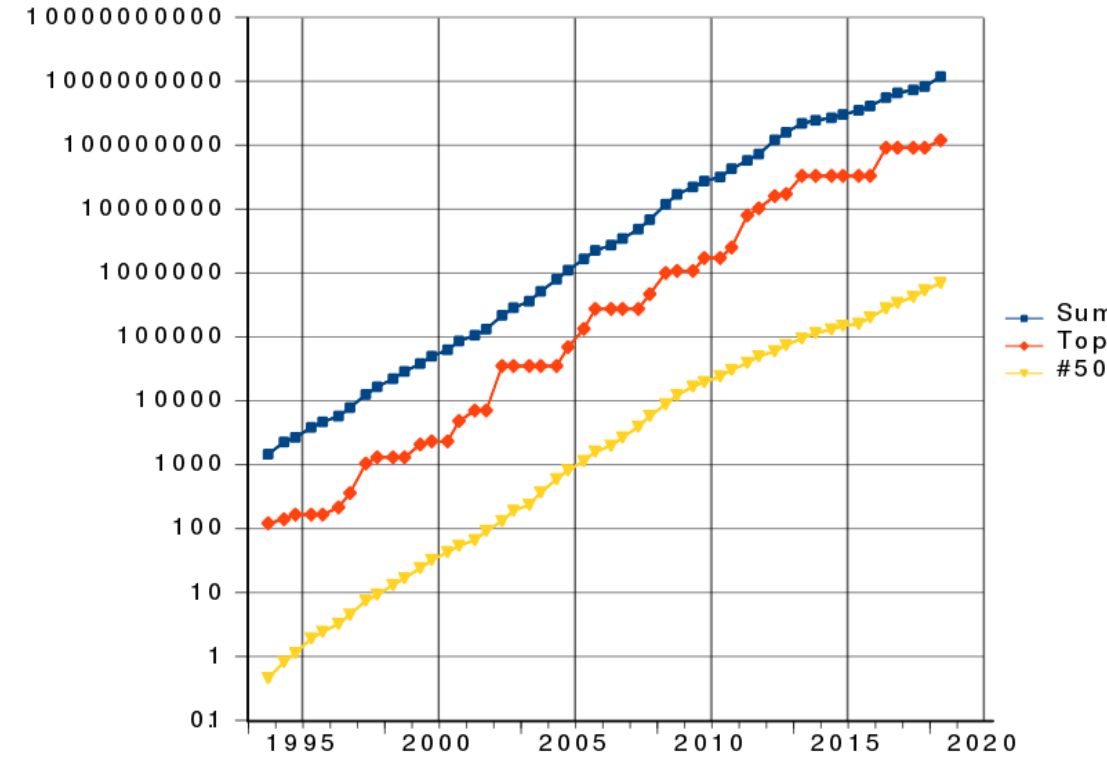# Performance optimization

Bootcamp for SahasraT

8th September 2018

Aditya Krishna Swamy

adityaks@iisc.ac.in
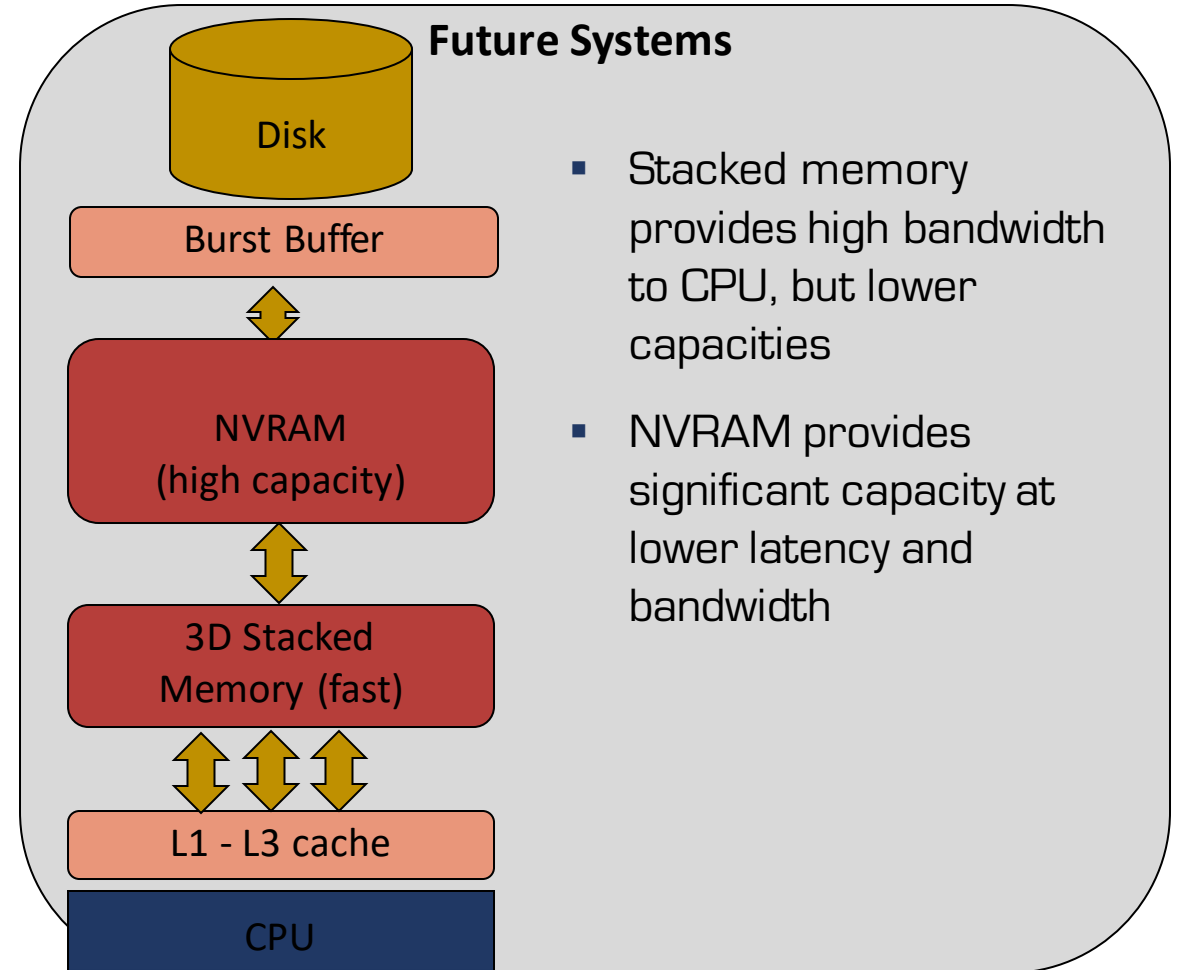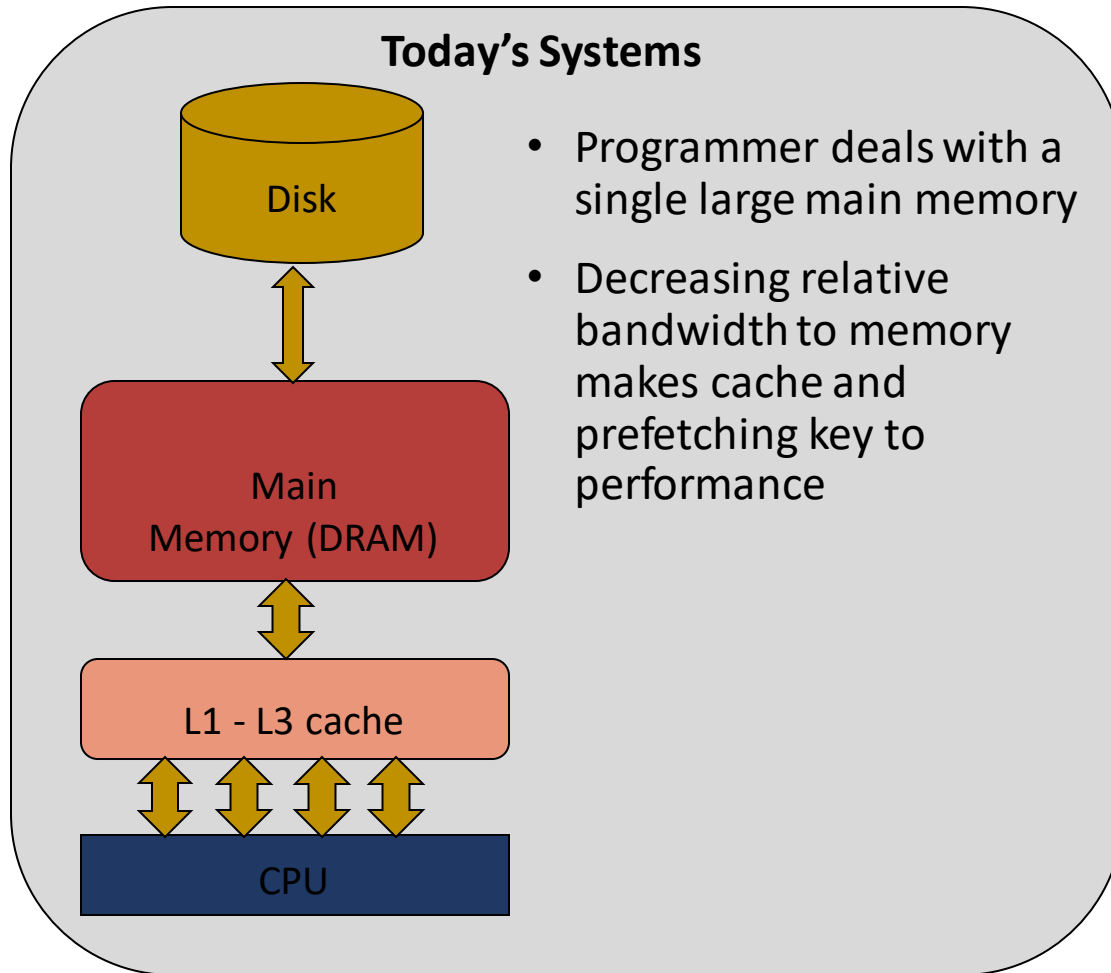
SERC, IISc

What is performance optimization

- A working code execution consumes resources – CPU, memory, power...

- Optmization - efficient use of these resources for faster results

- Available resources increasing rapidly

Scientific computing requires broad expertise
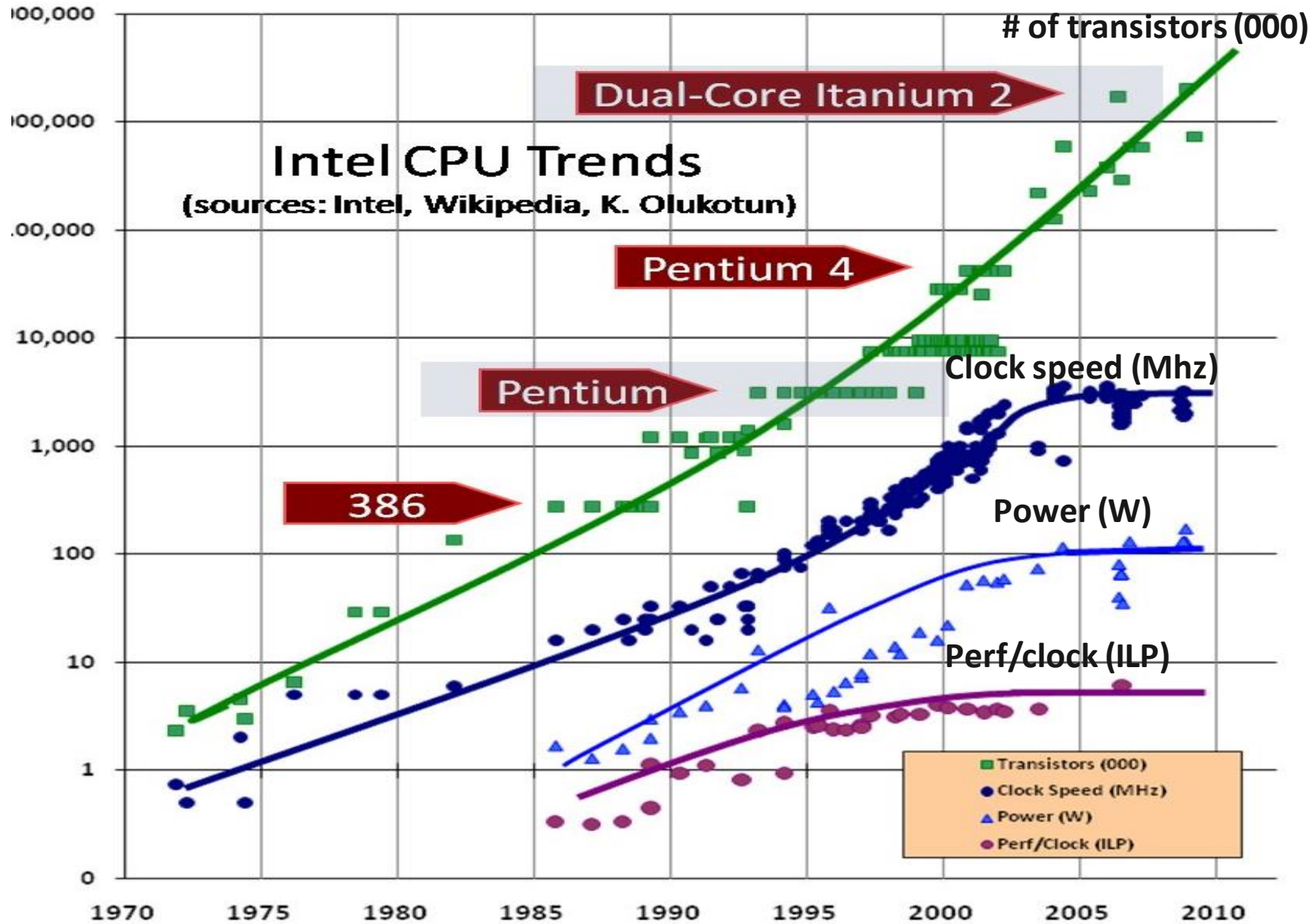*in addition to the research domain*
Knowledge of computer architectures, mathematical models
and numerical algorithms

Memory Systems are Getting Increasingly Complex

Today's Systems

Disk

Main Memory (DRAM)

L1 - L3 cache

CPU

- Programmer deals with a single large main memory
- Decreasing relative bandwidth to memory makes cache and prefetching key to performance

Future Systems

Disk

Burst Buffer

NVRAM (high capacity)

3D Stacked Memory (fast)

L1 - L3 cache

CPU

- Stacked memory provides high bandwidth to CPU, but lower capacities
- NVRAM provides significant capacity at lower latency and bandwidth

Future Challenges: Treat stacked memory as cache, or explicitly managed?
How can we take advantage of non-volatile nature of main memory?

The reason for disruption 10^{15} -> 10^{18} FLOPS FLOPS

Intel CPU Trends
(sources: Intel, Wikipedia, K. Olukotun)

# of transistors (000)

Dual-Core Itanium 2

Pentium 4

Pentium

386

Clock speed (Mhz)

Power (W)

Perf/clock (ILP)

- Transistors (000)
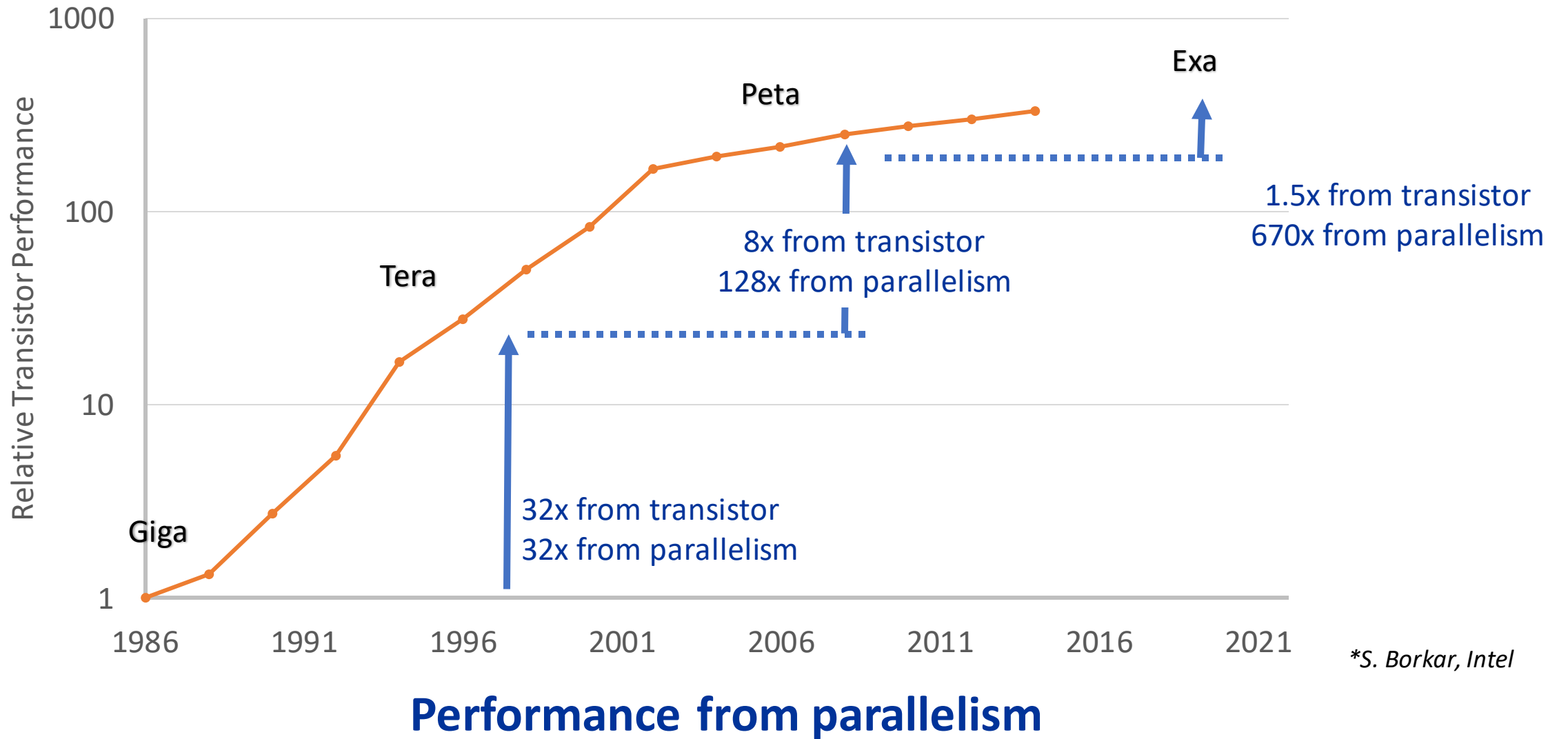- Clock Speed (MHz)
- Power (W)
- Perf/Clock (ILP)

Moore's Law still holds
Number of transistors
continues to increase

Dennard scaling does not:
Performance/Clock has flattened
out because
at 90nm transistor gates became
too thin to prevent current from
leaking out into the substrate

ILP: Instruction-Level Parallelism

From Giga to Exa, via Tera & Peta*

*S. Borkar, Intel

Performance from parallelism

Practical example