Trends in Technology

Integrated Circuit Logic

Transistor density increases by about 35% per year, quadrupling over four years. Combined with die size growth of 10-20% per year, transistor count grows 40-55% annually (doubling every 18-24 months)—known as Moore's Law.

Semiconductor DRAM

Capacity per DRAM chip has increased by 25-40% per year recently, doubling every 2-3 years. This growth rate has slowed over time, with concerns about whether it will stop mid-decade due to manufacturing challenges.

Semiconductor Flash

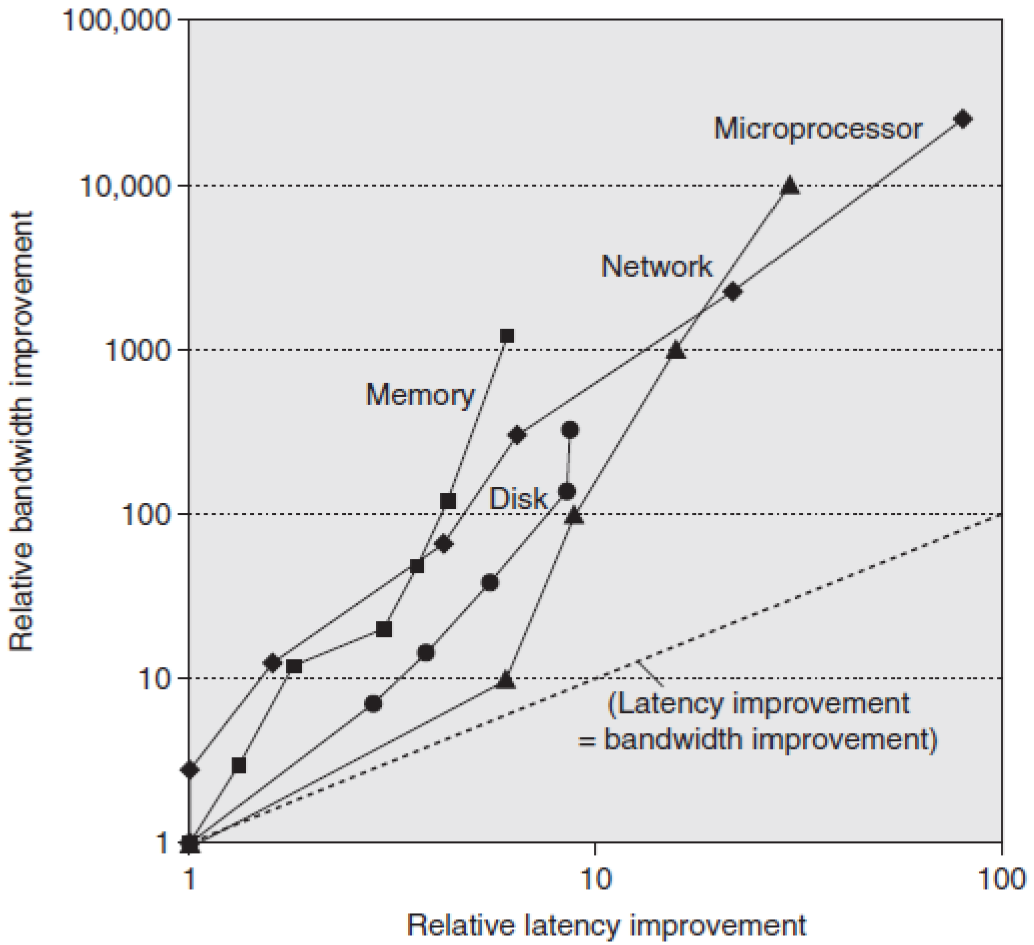
This nonvolatile memory has increased in capacity by 50-60% per year, doubling roughly every two years. In 2011, Flash memory was 15-20 times cheaper per bit than DRAM.

Magnetic Disk Technology

Density increased 30% annually before 1990, rose to 60% and then 100% by 1996, and dropped back to 40% after 2004. Disks are 15-25 times cheaper per bit than Flash and 300-500 times cheaper than DRAM.

Network technology, the fifth critical technology, depends on both switch performance and transmission system performance.

Bandwidth vs. Latency



- Bandwidth (throughput) is the total amount of work done in a given time, while latency (response time) is the time between start and completion of an event.
- As shown in the log-log plot on the left, bandwidth has dramatically outpaced latency improvements across technologies:
  - Microprocessors and networks: 10,000-25,000X bandwidth improvement vs. 30-80X latency improvement
  - Memory and disks: 300-1200X bandwidth improvement vs. 6-8X latency improvement
- A simple rule of thumb: bandwidth grows by at least the square of the improvement in latency. Computer designers should plan accordingly.

Performance Milestones

1

Microprocessors (1982-2010)

From Intel 80286 with 134,000 transistors to Core i7 with 1.17 billion transistors. Bandwidth improved from 2 MIPS to 50,000 MIPS (25,000X), while latency improved from 320ns to 4ns (80X).

2

Memory (1980-2010)

From 16-bit DRAM to 64-bit DDR3 SDRAM. Bandwidth improved from 13 MB/s to 16,000 MB/s (1,230X), while latency improved from 225ns to 37ns (6X).

3

Networks (1978-2010)

From 10 Mbits/sec Ethernet to 100 Gbits/sec Ethernet. Bandwidth improved 10,000X, while latency improved from 3000µs to 100µs (30X).

4

Hard Disks (1983-2010)

From 3600 RPM to 15,000 RPM drives. Bandwidth improved from 0.6 MB/s to 204 MB/s (340X), while latency improved from 48.3ms to 3.6ms (13X).

Scaling of Transistor

- Integrated circuit processes are characterized by feature size—the minimum size of a transistor or wire in either dimension. Feature sizes have decreased from 10 microns in 1971 to 32 nanometers in 2011, with 22nm chips under development.
- As feature size decreases:
  - Transistor density increases quadratically
  - Transistor performance improves approximately linearly
  - Operating voltage must be reduced to maintain reliability

Wire Delay Challenge

Wire Delay Problem

While transistors improve with decreased feature size, wires do not. Signal delay for a wire increases in proportion to the product of its resistance and capacitance.

Scaling Issues

As feature size shrinks, wires get shorter, but resistance and capacitance per unit length worsen. This relationship is complex, depending on process details, wire geometry, loading, and adjacency to other structures.

Design Impact

Wire delay has become a major design limitation for large integrated circuits and is often more critical than transistor switching delay. Larger fractions of the clock cycle are consumed by signal propagation delays.

While there are occasional process enhancements (like the introduction of copper) that provide one-time improvements, wire delay generally scales poorly compared to transistor performance.

Power Consumption

1 **Energy and Power Consumption**

System designers must balance performance with power consumption. Energy efficiency has become a primary design constraint across all computing segments.

2 **Thermal Management**

Heat generated by power consumption must be removed. Cooling solutions add cost, size, weight, and noise to systems, and may limit where systems can be deployed.

3 **Power Delivery**

Power must be supplied and distributed throughout the system. This requires significant infrastructure including power supplies, voltage regulators, and extensive power distribution networks.

From a system architect's perspective, these three primary concerns shape the entire design process and increasingly constrain what's possible in modern computing systems.