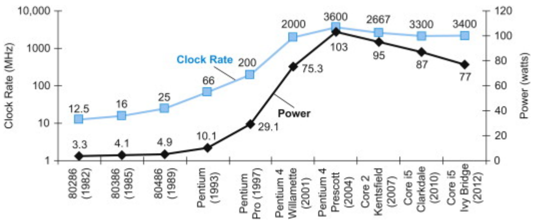
1.7 The power wall

The figure below shows the increase in clock rate and power of eight generations of Intel microprocessors over 30 years. Both clock rate and power increased rapidly for decades and then flattened off recently. The reason they grew together is that they are correlated, and the reason for their recent slowing is that we have run into the practical power limit for cooling commodity microprocessors.

PARTICIPATION ACTIVITY 1.7.1: Clock rate and power for Intel x86 microprocessors over eight generations and 30 years (COD Figure 1.16).

Start □ 2x speed



PARTICIPATION ACTIVITY 1.7.2: Clock rate and power trends.

1) As clock rates increased in early Intel processors, power _____.

○ increased

○ decreased

2) The _____ Intel processor consumed the most power.

○ 2004

○ 2010

3) Clock rates stopped increasing around 2004 because of power.

○ True

○ False

Although power provides a limit to what we can cool, in the post-PC era the really valuable resource is energy. Battery life can trump performance in the personal mobile device, and the architects of warehouse scale computers try to reduce the costs of powering and cooling 100,000 servers as the costs are high at this scale. Just as measuring time in seconds is a safer evaluation of program performance than a rate like MIPS (see COD Section 1.10 (Fallacies and pitfalls)), the energy metric joules is a better measure than a power rate like watts, which is just joules/second.

The dominant technology for integrated circuits is called CMOS (complementary metal oxide semiconductor). For CMOS, the primary source of energy consumption is so-called dynamic energy—that is, energy that is consumed when transistors switch states from 0 to 1 and vice versa. The dynamic energy depends on the capacitive loading of each transistor and the voltage applied:

$$\mathit{Energy} \propto \mathit{Capacitive\ load} \times \mathit{Voltage}^2$$

This equation is the energy of a pulse during the logic transition of 0 → 1 → 0 or 1 → 0 → 1. The energy of a single transition is then

$$\mathit{Energy} \propto 1/2 \times \mathit{Capacitive\ load} \times \mathit{Voltage}^2$$

The power required per transistor is just the product of energy of a transition and the frequency of transitions:

$$\mathit{Power} \propto 1/2 \times \mathit{Capacitive\ load} \times \mathit{Voltage}^2 \times \mathit{Frequency\ switched}$$

Frequency switched is a function of the clock rate. The capacitive load per transistor is a function of both the number of transistors connected to an output (called the ***fanout***) and the technology, which determines the capacitance of both wires and transistors.

With regard to the above figure, how could clock rates grow by a factor of 1000 while power increased by only a factor of 30? Energy and thus power can be reduced by lowering the voltage, which occurred with each new generation of technology, and power is a function of the voltage squared. Typically, the voltage was reduced about 15% per generation. In 20 years, voltages have gone from 5 V to 1 V, which is why the increase in power is only 30 times.

PARTICIPATION ACTIVITY 1.7.3: Energy and power.

1) Voltage = 4 V, frequency = 1 GHz, and power = 3 W. Frequency is increased to 6 GHz. What is the new power?

☐ W

Check Show answer

2) Voltage = 4 V, frequency = 1 GHz, and power = 3 W. Voltage is decreased to 2

V. What is the new power? Type as: 0.##

□ W

Check Show answer

Example 1.7.1: Relative power.

Suppose we developed a new, simpler processor that has 85% of the capacitive load of the more comp
processor. Further, assume that it can adjust voltage so that it can reduce voltage 15% compared to pr
B, which results in a 15% shrink in frequency. What is the impact on dynamic power?

Answer

$$\frac{\text{Power}_{new}}{\text{Power}_{old}} = \frac{\langle\text{Capacitive load} \times 0.85\rangle \times \langle\text{Voltage} \times 0.85\rangle^2 \times \langle\text{Frequency switche}}{\text{Capacitive load} \times \text{Voltage}^2 \times \text{Frequency switched}}$$

Thus the power ratio is

$$0.85^4 = 0.52$$

Hence, the new processor uses about half the power of the old processor.

The modern problem is that further lowering of the voltage appears to make the transistors too leaky, like water faucets that cannot be completely shut off. Even today about 40% of the power consumption in server chips is due to leakage. If transistors started leaking more, the whole process could become unwieldy.

To try to address the power problem, designers have already attached large devices to increase cooling, and they turn off parts of the chip that are not used in a given clock cycle. Although there are many more expensive ways to cool chips and thereby raise their power to, say, 300 watts, these techniques are generally too costly for personal computers and even servers, not to mention personal mobile devices.

Since computer designers slammed into a power wall, they needed a new way forward. They chose a different path from the way they designed microprocessors for their first 30 years.

Elaboration

Although dynamic energy is the primary source of energy consumption in CMOS, static energy consumption occurs because of leakage current that flows even when a transistor is off. In servers, leakage is typically responsible for 40% of the energy consumption. Thus, increasing the number of transistors increases power dissipation, even if the transistors are always off. A variety of design techniques and technology innovations are being deployed to control leakage, but it's hard to lower voltage further.

Elaboration

Power is a challenge for integrated circuits for two reasons. First, power must be brought in and distributed around the chip; modern microprocessors use hundreds of pins just for power and ground! Similarly, multiple levels of chip interconnect are used solely for power and ground distribution to portions of the chip. Second, power is dissipated as heat and must be removed. Server chips can burn more than 100 watts, and cooling the chip and the surrounding system is a major expense in warehouse scale computers (see COD Chapter 6 (Parallel Processors from Client to Cloud)).

PARTICIPATION ACTIVITY 1.7.4: Power wall.

1) Processor$_A$ has 75% of the capacitive load of processor$_B$. Processor$_A$ also has a 20% voltage reduction and 10% shrink in frequency. What is the relative impact on dynamic power?

○ $0.75 \times 0.80^2 \times 0.90 = 0.432$

○ $0.75 \times 0.20^2 \times 0.10 = 0.003$

○ $0.75 \times 0.80^2 \times 0.90 = 0.432\ \text{V}$

2) Which improvement has a bigger impact on power?

○ 25% reduction in voltage

○ 25% reduction in frequency switching

3) In the past 20 years, voltages have decreased from 5 V to 1 V. Why don't manufacturers continue to lower voltages to reduce power consumption?

○ Further lowering of voltage results in transistor leakage.

○ Voltage has no impact on power

4) Over the past 30 years, processor frequencies have continued to increase.

○ True

○

False

Provide feedback on this section