

Machine Organization and Rent's Rule

E.F. Rent of IBM published two internal memoranda in 1960 that contained the log plots of "number of pins" versus "number of circuits" in a logic design [9.26]. These data tend to form a straight line in a log-log plot and yield the relationship

$$N_p = K_p N_g^b \quad (9.1)$$

Here, N_p is the number of pins or the number of external signal connections to a logic block, N_g is the number of logic gates in the block, b is the Rent's constant, and K_p is a proportionality constant. The values of K_p and b for the IBM computers were reported to be 2.5 and 0.6, respectively.

Rent's rule was later applied by IBM researchers to derive the average interconnection length in a gate array as a function of number of gates and gate pitch [9.27]-[9.32]. This was obtained by hierarchically dividing the logic design into quadrants and calculating the number of connections to a block from Rent's rule. By setting the upper limit on wire length equal to the size of the block that forms the next level of the hierarchy, a series sum of interconnection lengths was formed that establishes an upper bound on total wire length.

Because Rent's rule is an empirical result obtained by observing existing designs, it is useful in predicting the pin requirements and average interconnection lengths of well-studied architectures and follow-on computers that have designs similar to current systems, but it may give misleading results if extended to dissimilar architectures and too far into the future. The limitations of Rent's rule must be recognized. For accurate predictions, it is necessary to understand the design from which the initial Rent's data were obtained and to ensure that the architecture and implementation of that design are similar to the system being studied. Like any empirical (or even fundamental) relationship, Rent's rule can be misleading if it is applied without adequate understanding.

The key factors that affect Rent's constants include machine architecture, organization, and implementation. The design philosophy and methodology also affect Rent's constants. If the machine from which the initial data were obtained and the computer design for which the predictions are sought have similar machine organizations and implementations, the model will be successful. On the other hand, if the predictions are made for a system with an entirely different design philosophy from the one from which Rent's data were obtained, the results will have little meaning.

Many architectural and implementation factors affect the pin count and interconnection requirements and, as a result, the Rent's constant. For example, to ensure the highest possible speed, multiplexed I/O pins and serial transmission of data are avoided, and this results in higher pin counts. On the other hand, to achieve low-cost packaging in commercial microprocessors and memories,

bidirectional and multiplexed I/O pins and partially serial data transformation are employed.

It is important to note that two chips with exactly the same number of gate counts may have different pin counts due to package cost considerations. As a result, drawing conclusions concerning on-chip interconnection lengths from pin count data may be misleading because two chips with identical lay-outs and on-chip interconnection lengths may have drastically different pin counts if multiplexed or bidirectional pins and serial I/O ports are employed in one of the chips. For example, chip vendors usually offer a version of their 32-bit microprocessors that is basically identical to the original one except the external interface is 16 bits. The concerns are similar when the average interconnection calculations derived for a design methodology (for example, gate arrays) are applied to a chip designed by another methodology (for example, custom-designed microprocessors).

Another design area that illustrates the relation between average interconnection length and architectural design choices is pipelining. As described in Section 9.3, throughput of a design can be improved by pipelining, but this also increases the communication requirements (and therefore wiring) between the subunits because of the dependencies between the instructions that are executed simultaneously.

TABLE 9.5 Rent's Constants for Various System Types

System or Chip Type	Exponent β	Multiplier K_P
Static memory	0.12	6
Microprocessor	0.45	0.82
Gate array	0.50	1.9
High-speed computer		
Chip and module level	0.63	1.4
Board and system level	0.25	82

The following considerations relate to supercomputers and mainframes:

- Because of their high level of complexity and the limited design times between new generations and models. Supercomputers and mainframes are usually implemented in gate arrays.
- These high-performance systems require a parallel I/O and avoid multiplexed or bidirectional pins.

- High performance also requires concurrency (pipelining and parallelism), which increases the communication demands among the subsections of the CPU. As a result, even if the chips are implemented as custom designs, their interconnection requirements may still be comparable to gate arrays.

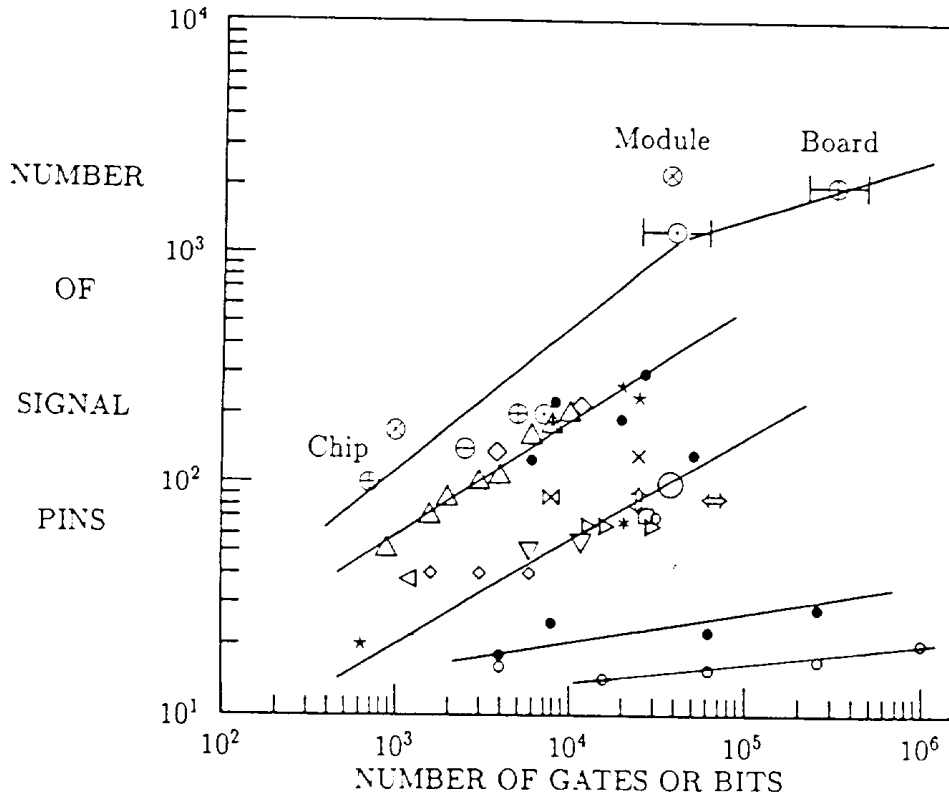
Trends in high-performance systems are important for lower-cost designs because, historically, advanced design techniques such as pipelining were first implemented in mainframes and eventually found their way into microprocessors.

This chapter considers a wide variety of systems. Rent's data obtained from super- and mainframe computers and several commercial integrated circuits will be used to predict the pin requirements and average interconnection lengths of different architectures and implementations. Figure 9.11 shows the pin counts of several chips and packages, and Fig 9.12 plots the same data classified according to the type of system instead of product identification. The data are grouped into distinctive classes, and the slopes of the lines in these plots determine Rent's constant β in Eq. 9.1 for the listed system types.

Expectedly, memory devices have small pin counts. Faster static RAMs have higher pin counts than DRAMs because SRAMs usually have separate pins for all address bits. In DRAM chips, half the address (row address) is sent during one clock cycle and the other half during the following cycle (column address), and they share the same pins.

Commercial microprocessors form the next group. In microprocessors, data input, data output, and occasionally address share the same pins. In addition, microprocessors are self-contained systems that constitute a functionally independent portion of a larger system. This functional completeness reduces the pin requirements.

In gate arrays, the placement of the gates and circuit design is done long before the logic is designed. In this way, most of the semiconductor processing can be done ahead of time, and only the metal wiring steps are left to personalize the design. Processed wafers are stockpiled and quickly turned around after chip definition. Much of the development cost is shared among many designs, reducing the cost. This makes them very attractive for low-volume parts. Because of the rigidity of placement, however, gate arrays are not as efficient as custom designs. Because they have to meet the requirements of a wide range of designs. For any given design, gate arrays are not as dense and fast as a custom implementation would be. In addition, gate arrays require more wiring space because the placement is done before the logic design. Wiring is more efficient in custom designs because placement can be optimized to minimize interconnection lengths. In the pin count plot, gate arrays are immediately above microprocessors because, due to their limited integration density, gate arrays are not as functionally self-contained as microprocessors and, as such, require more pins.



High performance computers	Microprocessors
⊕ IBM ECL gate array	★ Intel 8008
- ⊙ - IBM 3081 TCM	◊ Intel 8080,8085,8086
- ⊕ - IBM 3081 board	▷ Intel iAPX-43xxx
⊗ NEC SX	□ Intel 80286
Gate Arrays	• Intel 80386
△ LSI logic CMOS	◁ Motorola 6800
• Toshiba CMOS	• Motorola 68000
★ Fujitsu CMOS	○ Motorola 68020
◊ Hitachi CMOS	▽ Zilog Z8000
⊙ NTT ECL	× Fairchild Clipper
⊖ Siemens ECL	◦ μVAX 32720
Memory Chips	↕ Bellmac-32A
• Static RAM	↔ HP 32bit CPU
◦ Dynamic RAM	× Stanford MIPS
	▽ Berkeley RISC1

FIGURE 9.11 Rent's curves for various digital systems. Data points are classified according to product identification.

High-performance computers have the largest pin counts because they use more parallelism and pipelining, which in turn, require a greater number of simultaneous information transfers. They normally avoid multiplexed and bidirectional I/O when possible. Module and board levels are also shown on the plot as if they were individual chips. This symbolizes a design in which one single module or board is mapped to a chip with an identical logic structure and the same pin count. When

the unit (chip, module, or board) contains the totality of a CPU, the rate of increase in pin count drops; however, this does not occur until the late stages of completion. In the JBM 3081, one board contains a CPU, and each board consists of nine modules: as a result a module contains one-ninth of a CPU but still appears on the same Rent curve as the chips. When the gate count of the unit increases and contains more of the entire CPU, the pin count will move to another curve with a less steep slope (smaller Rent's constant) as illustrated in Figs. 9.11 and 9.12.

Rent's constants obtained from Fig. 9.11 are listed in Table 9.5. The quantitative results support the earlier observation that these constants are determined by the architecture and implementation of the system and that pin and interconnection requirements are more demanding in high-performance systems with large amounts of parallelism and pipelining.

The model described in the remainder of this chapter is very flexible; the architecture-dependent aspects can be changed by using a different Rent constant, and the results can be compared.

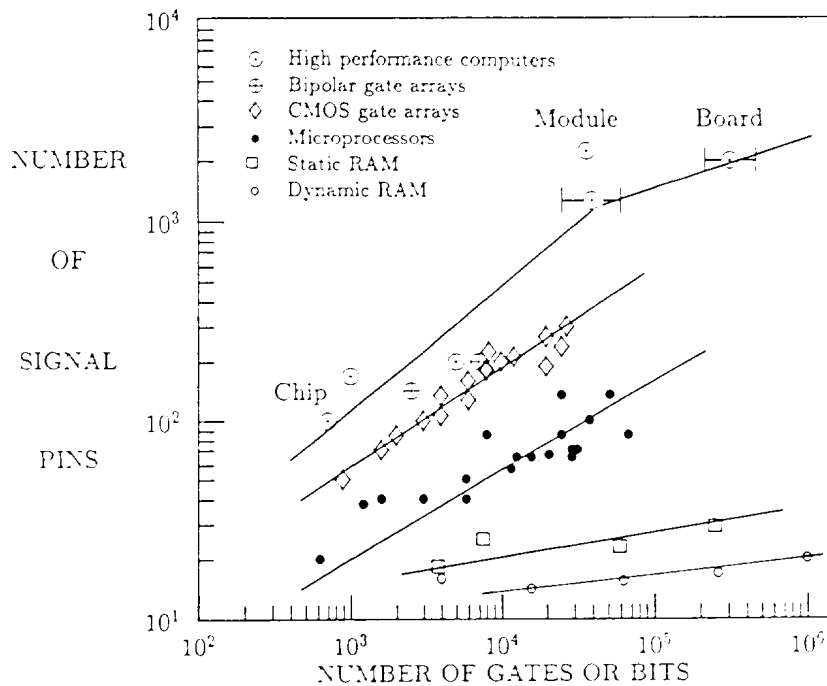


FIGURE 9.12 Rent's curves according to system and chip types