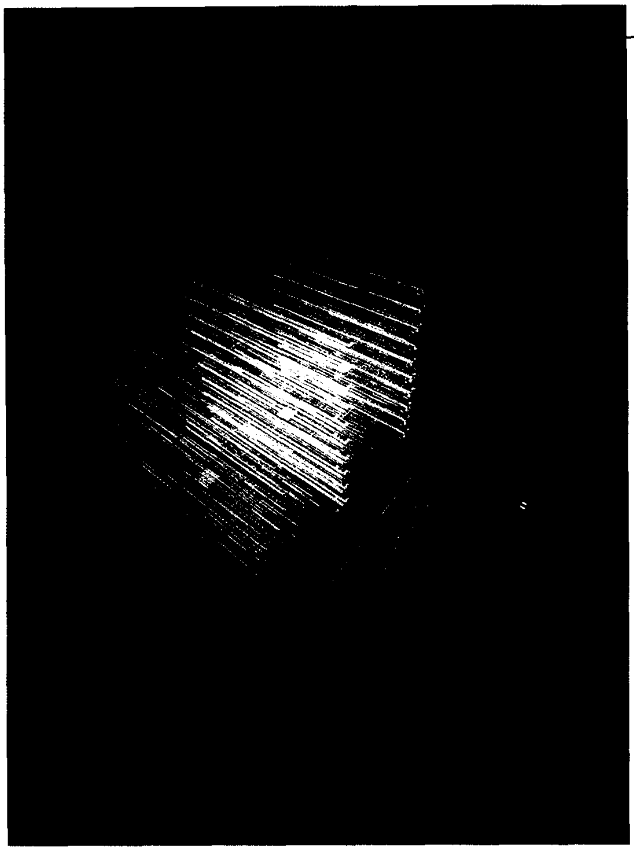


The Stock Market © 92 Steve Grohe

Looking at the materials and thermal alternatives for scaled, next-century VLSI/ULSI interconnects

by Soo-Young Oh and Keh-Jeng Chang

2001 Needs for Multi-Level Interconnect Technology

Thanks to advanced scaling techniques over the past 20 years, device performance and operating speeds have skyrocketed. Clock frequencies already exceed 200 MHz in submicron RISC microprocessors. As the minimum feature size has continued to scale down to submicron proportions, however, minimum interconnect line widths and spacings have, of necessity, followed. As a result, interconnect performance has become a limiting factor impinging on circuit performance. The RC delay of lines increase, and tend to limit the length of global routing. Crosstalk becomes a problem, and limits the scaling of metal pitches. The current densities also increase, approaching the electromigration limit. Ultimately, this class of performance bottlenecks transcend good circuit design. Clearly, improving interconnect performance is the key to reducing and controlling degradation within acceptable levels.

Technology in memories, for example, has proceeded to the point where a new generation of SRAMs are expected every 2-3 years [1]. Their capacity will be in the

8755-3996/95/$4.00©1995IEEE

range of 256 Mb to 1 Gb in 2001. The minimum feature size will be scaled to 0.18-0.20 μm. The interconnect width and spacing, which is usually twice that of the minimum feature size, will be reduced to 0.35-0.4 μm. In these deep sub-half-micron technologies, IC performance shortcomings will be unacceptable, rendering the device unusable. Without doubt, better multi-level interconnect models will be required for next-century ICs.

We need only look back at the history of the interconnect issue to see how far we have come, and where we need to go. Table 1 shows the effects of scaling on FET transistors and interconnects with the ideal scaling proposed by Dennard [2] in 1974. Given that the resistance of transistors (R_{tr}) stays the same for a given scale factor, the gate capacitance (C_{gate}) of the next stage is reduced by 1/S, where S is the scaling factor of the minimum feature size and is larger than 1. The gate delay ($R_{tr}*C_{gate}$) is thus reduced by 1/S, and so the device speed is improved, as expected. However, for global interconnect lines, the line resistance (R_{int} (global)) increases by S^2*Sc, where Sc is the scaling factor of the chip size increase, and is larger than 1; and where the S^2 comes into consideration because the line thickness and width are reduced by 1/S. On the other hand, the line capacitance is only increased by Sc, because the line capacitance per unit length stays the same and the line length increases by Sc. The RC time constant of the global line is thus increased by S^2*Sc^2.

Both S and Sc are approximately the same value for each generation. Therefore, the delay of the global line is increased by S^4 with ideal scaling. In practice, ideal scaling is difficult to implement rigorously. Quasi-ideal scaling has been proposed and is followed more or less today by the IC industry. In quasi-ideal scaling, the vertical dimension of interconnect (e.g., line thickness and dielectric thickness) are scaled by the square root of the lateral scaling factor (S). Still, the delay of the global line is increased by S^3. This clearly illustrates why the interconnect delay is so drastically increased with scaling and becomes a problem.

In practice, ideal scaling is difficult to implement

For example, the interconnect delay of a 10 mm line is about 1 nsec when the line width and spacing are 0.9 μm. However, it will increase to 6 nsec when the line width and spacing are scaled to 0.3 μm. For the planned 0.3 μm generation of devices, the clock frequency of RISC microprocessors is expected to be 500 MHz-1 GHz. The clock period is 1-2 nsec. In these circuits, an interconnect delay of 6 nsec will not be acceptable. The current density will also increased by S in the ideal scaling and by the square root of S in quasi-ideal scaling. Thus, the reliability of electromigration will also degraded with scaling. When the minimum width and spacing reaches below sub-half micron, the aspect ratio of the line will be increased, and the crosstalk will be very significant and limit the interconnect routing very severely. As explained previously, interconnect degradations with scaling are significant and cannot be handled separately by the device technology, circuit, or system design. Concurrent and global optimization of the interconnect system is therefore necessary.

Table 1
Effects of Scaling on Interconnect

	Ideal Scaling	Qusi-Ideal Scaling
R_{tr}	1	1
C_g	1/S	1/S
R_{int} (global)	S^2Sc	$S^{3/2}$ Sc
C_{int} (global)	Sc	$S^{-1/2}$ Sc
R_{int} C_{int} (global)	S^2 Sc^2	S Sc^2
J_{int}	S	$S^{1/2}$

Interconnect Modeling

In the very lossy on-chip interconnect, the inductive voltage drop is negligible compared to the resistive voltage drop up at clock frequencies of 1- 2 GHz. Thus, the on-chip interconnect line may usually be approximated as an RC line. Traditionally, Ohm's law is used to calculate line resistance, and the parallel-plate capacitor model to calculate line capacitance. When the aspect ratio between the line width and spacing is increased, the fringing capacitance also becomes significant. Resistance and capacitance are usually calculated using analytical equations based on conformal mapping, as with microwave circuits.

However, when the line pitch is decreased and the topology of the on-chip interconnect becomes complicated, these traditional methods are not acceptable. In order to model and analyze interconnect lines, 2-D and even 3-D electrostatic field simulations are necessary, especially calculating capacitance accurately in the submicron interconnect technology. Several closed-form equations have been derived or approximated from measured data for fast capacitance calculations, but their application to general submicron interconnect is rather limited, especially for the coupling capacitances. Numerical techniques have also been developed for rigorous interconnect capacitance extractions. They fall into the following three categories: (1) finite-difference method, (2) finite-element method, (3) Green's function method. Although they are very accurate, ad hoc executions of those programs to calculate interconnect parameters for VLSI/ULSI design and analysis are too time-consuming to be practical.

The tool used in our study to model our hypothetical interconnect system, which investigated various materials and thermal alternatives, was Hewlett Packard's HIVE (HP's Interconnect Value Extractor). With HIVE, parameter data and design curves are interpolated from the look-up table generated by the batch-mode simulation of 2-D (FAP2) and 3-D (FCAP3) electrostatic field solvers. HIVE consists of the batch-mode CapSim, and interactive-mode CurveGen. CapSim is run automatically when a new interconnect geometry is available or modfied. In this mode, 2-D and 3-D capacitance simulations for the representative interconnect structures are executed to call FCAP2 and FCAP3 as needed.

The resulting values are stored in CapFile. The information contained therein is used with CurveGen, which matches the specific structure with CapFile, or interpolates or extrapolates the CapFile data to generate interconnect design curves.

Interconnect Alternatives

In order to solve these interconnect degradations, several alternatives have been proposed in the technology, circuit, and system. They are: (1) change the metallurgy of lines from aluminum (Al) to copper (Cu) to reduce the line resistance. (2) place non-scaled Al lines on the additional metal layers on top of the minimum-pitch layers for a long-range global interconnect lines. (3) replace the interlevel dielectric with low permitivity material to reduce the line capacitance. (4) insert repeaters in the long interconnect lines in circuit design. (5) operate circuits at low temperature (77 K) to reduce the line resistance.

In order to evaluate these approaches, the relationships between all factors leading to degradation, versus the anticipated improvements afforded by factors (1) through (5) above, have been analyzed. First, we adopt a scaling technique.

Ideal scaling will degrade the interconnect delay too much, whereas only lateral scaling will make the crosstalk intolerable.

Quasi-ideal scaling is more or less followed by the IC industry

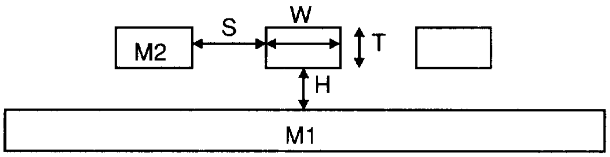
Quasi-ideal scaling is more reasonable in this regard and thus has been adopted by the industry. Therefore, we apply it in our evaluation. We also assume that the interconnect line width and spacing of the current CMOS technology are 0.9 μm; and the metal line thickness and dielectric thickness between the metal layers are 0.7 μm.

Now we will consider three future generations of scaled interconnect technologies, with line widths and spacings at 0.7 μm, 0.5 μm and 0.3 μm. We also assume a minimum improvement between generations of 30 percent. Table 2 lists the geometry and other important parameters of each generation of, in this case, scaled CMOS processes.

With this interconnect geometry scaling, the capacitances and resistances per unit length of each generation have been calculated (Table 3). The non-scaled 2-μm wide Al line case is also calculated. The line-to ground capacitance (C_g) decreases by half from the 2-μm wide line to the 0.3-μm wide line, whereas the 2-side coupling capacitance (C_c) increases about twice. The total line capacitance only increases by 20 percent in this scaling. However, the line resistance increases more than 16 times. Thus, the main factor for RC delay increase with scaling is due to the resistance increase.

Table 2
Geometry of Analyzed IC Processes

	Non-Scaled 2 μm line	CMOS1	CMOS2	CMOS3	CMOS4
W (μm)	2.0	0.9	0.7	0.5	0.3
S (μm)	2.0	0.9	0.7	0.5	0.3
T (μm)	1.0	0.7	0.62	0.52	0.4
H (μm)	1.5	0.7	0.62	0.52	0.4
V_{dd} (volts)		5.0	5.0	3.3	3.3
Freq (MHz)		50.0	65.0	84.5	109.9



Now we calculate the capacitances are for the cases of low permitivity polyimide dielectric. As shown, capacitances are reduced by 33 percent, which is the same as the permitivity reduction. The resistances of the low temperature and Cu cases are reduced by 40 percent in Cu and 12 times in low temperature, compared with Al lines at room temperature.

Three figure-of-merits are selected to compare and evaluate alternatives (1) to (5); interconnect line delay, crosstalk noise, and the maximum current density for the electromigration reliability. In order to calculate the interconnect line delays, a modification of Sakurai's distributed RC delay model [2] is used. RISC microprocessors run at around 50 MHz in the current CMOS technology. The critical path usually consists of about 10 logic gates in RISC microprocessors. In the optimum design, the gate delay and interconnect delay are balanced. The acceptable interconnect delay (T_{crit}) is about 1 nsec (1/20 of cycle time). The maximum clock frequency of each generation is assumed to increase a minimum of 30 percent. The maximum line lengths (L_{max} (delay)) before line delay reaches T_{crit} are in Fig. 1 for the minimum pitch Al lines, Cu lines, and Al in 77 K.

L_{max} (delay) of the non-scaled 2-μm width/spacing lines on the additional layers on top of the minimum-pitch layers have also been calculated and plotted. Odd-mode switching is assumed (adjacent signal lines are switching to the opposite direction). As seem L_{max} (delay) of the minimum pitch Al lines is degraded below 2 mm in the 0.3 μm case. Even though the increase of the clock frequency is conservatively estimated, the degradation of L_{max} (delay) is unacceptable for Al lines for the global routing (> 10 mm). The Cu lines improves L_{max} (delay) marginally due to the 40 percent improvement of resistivity, but it is still not enough for the global routing below 0.5 μm case.

Due to the significant reduction of resistance, low-temperature Al lines and non-scaled Al lines improve L_{max}(delay) above 10 mm even in the 0.3 μm case. Low-permitivity interlevel dielectrics have been simulated using polyimide (permitivity = 2.6) for the above four cases. Polyimide has 33 percent lower permitivity and improves L_{max} (delay) by 20-30 percent.

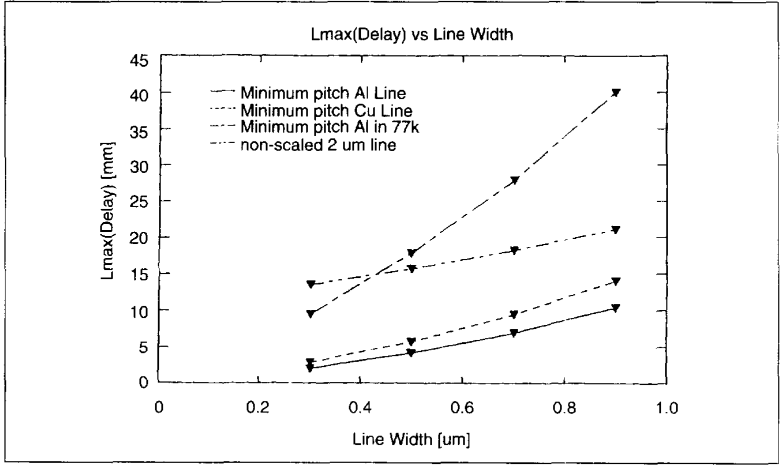
When it comes to crosstalk noise, three parallel lines driven by buffers were simulated by SPICE with the center line as the victim line. The first and third lines were switched simultaneously in the same direction, while the center line was kept constant. The crosstalk noise was coupled to the center victim line. The maximum allowable noise is set to the 20 percent of the power supply (V_{dd}), which is close to the threshold voltage (V_t) of a MOSFET transistor.

Changes in metallurgy and dielectrics will improve interconnects

If the noise is above V_t, the charge stored in the dynamic node will leak. Maximum line length (L_{max} (crosstalk)) have been calculated and plotted in Fig. 2. Only the non-scaled lines are acceptable for global routing. The other three approaches have very small L_{max}(crosstalk) in the sub-half-micron range. In crosstalk noise, the primary factor is the interline coupling capacitance, and the effect of the line resistance is minimal or secondary. The capacitance of the other approaches remain the same, and the improvement of the line resistance in Cu or low temperature is not appreciable. The above cases with polyimide as an interlevel dielectric have also been simulated. Polyimide improves L_{max}(cross-talk) by 25-30 percent.

Table 3
Calculated Capacitances and Resistances for All Alternatives

	Non-Scaled 2 µm line	CMOS1	CMOS2	CMOS3	CMOS4
C_g (aF/µm)	83.3	80.6	71.7	62.0	49.1
C_c (aF/µm)	80.0	99.8	110.0	124.5	150.2
C_{unit} (aF/µm)	163.3	180.4	181.7	186.5	199.3
C_{unit} (aF/µm): PI	108.8	120.1	120.0	124.2	132.7
R (ohm/µm): A1	1.5e-2	4.7e-2	6.9e-2	1.15e-1	2.5e-1
R (ohm/µm): Cu		2.8e-2	4.1e-2	6.9e-2	1.5e-1
R (ohm/µm): LT		4.0e-3	5.8e-3	9.8e-3	2.1e-2



1. L_{max}(delay) vs. line width for minimum-pitch Al lines, Cu lines and Al lines at 77 K.

As discussed above, the scaling increases the current density in interconnect lines and degrades the electromigration reliability. In the uni-directional lines such as V_{dd} or ground, the increase of current density can be alleviated just by widening the line width. However, the current density in the bi-directional lines such as the signal lines can't be reduced when the line width is widened. Because the line capacitance increases, the rate of discharge is not decreased with the wide line. Therefore, the bi-directional lines will limit the electromigration reliability. The average current in bi-directional signal line at the driver can be expressed as follows, and must be within the limit set by the electromigration reliability:

$$I = 2(C_{int} + C_{gate})V_{dd}\,F_{max} < J_{max}W\,t$$

where V_{dd} is the power supply voltage, and F_{max} is the maximum clock frequency. W and t are the line width, and thickness, respectively. J_{max} is the limit of current density limit set by the electromigration reliability. For an Al line at room temperature, J_{max} is set to 9E5 A/cm^2. The maximum line lengths (L_{max}(E/M)) are calculated for the minimum pitch Al line and the non-scaled line and plotted in Figure 3. The Al line is not acceptable for the global routing below 0.5 µ. Only the non-scaled line is useful for the global routing due to the large cross-section. L_{max}(E/M) for the Cu line and low temperature Al line are not calculated because J_{max} is not available. But J_{max} for these cases should be larger by an order of magnitude. Thus, both approaches improve L_{max}(E/M) significantly and make them available for global routing. Calculations for polyimide shows it improves L_{max}(E/M) by 33 percent.

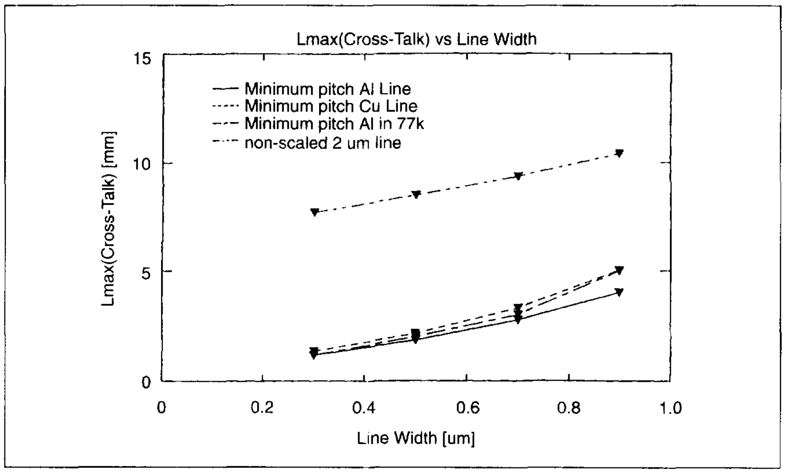
When the interconnect resistance and capacitance are larger than the repeater on-resistance, and gate capacitance, respectively, the repeater delay is negligible compared with the line delay. In this case, interconnect delay increases by the square of the interconnect length because both capacitance and resistance increase linearly with length.

The use of repeaters makes the line delay reduced by the same factor of the number of the repeaters. When the line is divided into k subsections using k repeaters, the line delay of each subsection is reduced by the square of k. There are k subsections and the total line delay is reduced by a factor of k. The repeater not only improves the interconnect delay, but also improves the cross-talk noise and electromigration reliability. By dividing a long line into smaller subsections with repeaters, the effective lengths for the cross-talk noise and bi-directional electromigration become the lengths of the subsections and reduced by a factor of k.

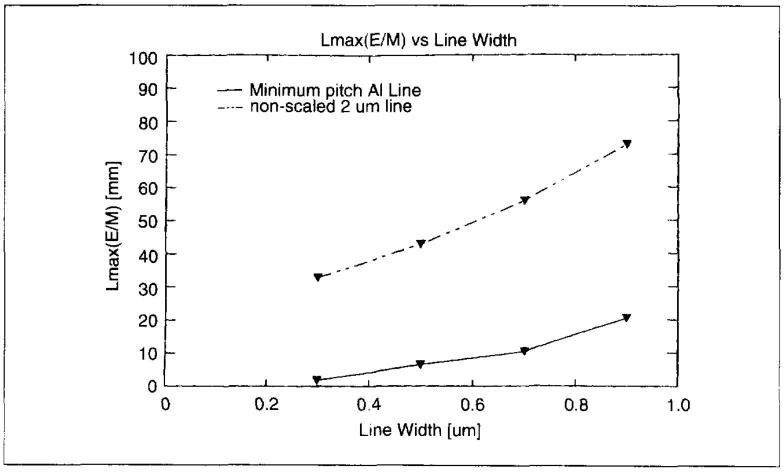
Other possibilities: non-scaled lines in higher metal layers, repeaters, low-T operation

Conclusion

The above evaluations show that Cu lines will improve the electromigration reliability, but the interconnect delay marginally, due to the small improvement of the resistivity. The crosstalk noise is not much improved, because the coupling capacitances basically remains the same. Furthermore, Cu is a new interconnect material for ICs, and much developmental work still needs to be done. Low temperature operation improves the interconnect delay and electromigration reliability, but it doesn't improve the crosstalk noise, and increases the cost of the system packaging. Low permitivity interlevel dielectrics will improve all three figure-of-merit ratings by 25-30 percent, and the repeater will improve all by the factor of the number of repeaters. The non-scaled lines in a higher metal layer can improve all three figure-of-merits significantly, and are thus acceptable for global routing. However, due to various factors, it may not be enough by itself. The optimum approach will therefore be the combination of additional layers of non-scaled lines, low permitivity interlevel dielectric, and the use of repeaters to maximize the performance, noise immunity, and reliability, and to minimize the cost.



2. L_{max} *(crosstalk) vs. line width for the minimum-pitch Al lines, Cu lines and Al lines, at 77K.*



3. L_{max}*(E/M) vs. line width for the minimum-pitch Al lines and non-scaled 2 μm lines at higher level.*

Beyond 2001, the needs for the interconnect technology are as follows. First, global and near-perfect planarization is needed for the large number of multiple metal layers to maintain the non-scaled lines. Planarization is already required due to the small depth-of-focus in the sub-half-micron lithography. Thus, it should be the first priority.

The low-permitivity interlevel dielectric will be the next priority. It is a fairly new material to MOS, but it has already been used in bipolar. It will take less time to be implemented than Cu and will have more return-on-investment. Implementing Cu will take longer. When it is ready, it will be more desirable for use in the higher metal layers as non-scaled lines in the multi-layer interconnect technology. If these alternatives are still not enough, low-temperature operation will be needed as a last resort.

Acknowledgement

We wish to thank Jung Lin, Kuang Chiu, Ken Lee, Yao-Feng Lin, and Kit Cham for an extraordinary amount of help and inspiration.

Soo-Young Oh and *Keh-Jeng Chang* are with Hewlett Packard, Palo Alto, CA.

References

1. MICRO TECH 2000 Workshop Report, National Advisory Committee on Semiconductors, August, 1991.

2. R.H. Dennard et al., "Design of ion implanted MOSFET's with Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits*, Vol. SC-9, pp. 256-268, Oct. 1974.

3. T. Sakurai, "Approximation of Wiring Delay on MOSFET LSI," IEEE Journal of Solid-State Circuits, pp. 418-426. August 1983.

IEEE Presents Its Plans for Conference Proceedings

THE 1995 IEEE PREPAID ORDER PLAN (POP)

Subscribe to the 1995 POP plan and you'll get about 130 major 1995 IEEE conference proceedings—for $11,995! This is the standard plan libraries have received for decades.

THE 1995 IEEE PREPAID ORDER LAN PLUS (POP)

Receive the 130 major conferences discussed above *plus* about 80 new 1995 IEEE conferences, workshops, and symposia in rapidly-emerging technologies—for $16,975! POP Plus is *all* the 1995 IEEE conferences, in electrical engineering and computing in one plan.

SUBSCRIBERS TO BOTH PLANS receive about a 38% savings off the total list prices of individual conferences.

For more information, contact Robert S. Hohwald, IEEE Tel: (908) 562-3999.

Brain Buster

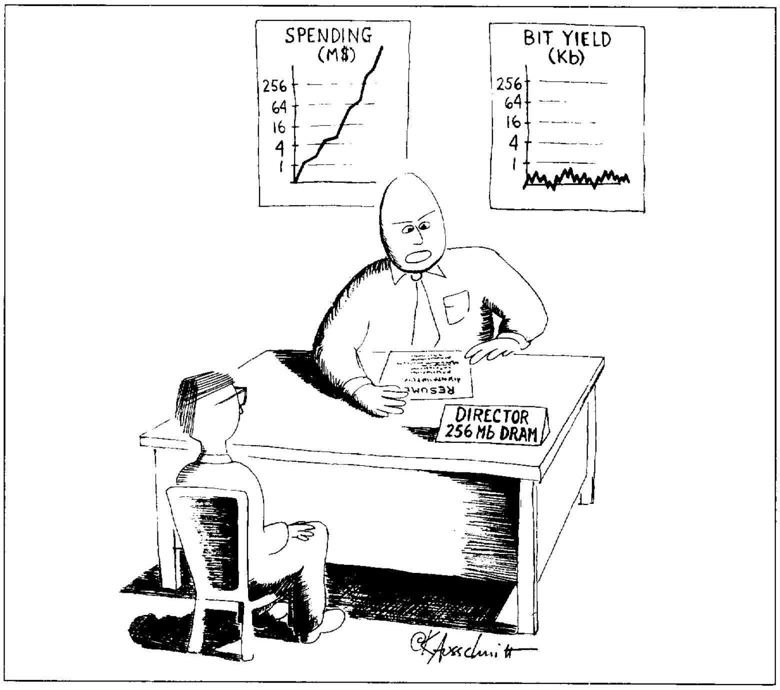
Here's an interesting gambling game. You pay the house a fee for the privilege of flipping a coin. The house will pay you a dollar for each flip it takes you to get the first head. What should the fee be to make it a fair game?

Answer to last issue's question:

The following table shows the moves that fufill the conditions to get everyone across the river:

Trip no.	Row over	left on near side	row back	left on far side
1	G W	K Q P PR	G	W
2	Q PR	K P G	W	Q PR
3	K P	G W	K Q	P PR
4	G W	K Q	P PR	G W
5	K P	Q PR	W	G K P
6	Q PR	W	G	K Q P PR
7	G W	no one	-	K Q P PR G W

From: *The Unofficial IEEE Brainbuster Gamebook*, by Donald Mack.



"I see here, Mr. ...er..., young man, you were able to design and fabricate a fully functional 256Mb DRAM for your senior science project—would you care to elaborate on that experience?"