The relative volumes of wires and components in a computer

Warren D. Smith
wds@research.NJ.NEC.COM

date = June 24, 1999

Abstract — **In a computer with a fixed number of "components" of fixed sizes and a fixed "wiring diagram," what wire thicknesses will result in maximum performance? For a model of neurons, Chklovskii and Stevens argued that the "wires" should optimally consume $x = 3/2$ times the volume of the "components." For 2D VLSI with thin sheet conductors we argue $x \to 0$ is best (albeit difficult to achieve in practice). For 2D or 3D VLSI with cylindrical wires, $x \to \infty$ is best if RC delays in wires dominate; $x \to 0$ is best if source resistances dominate. For fiber optics and superconducting transmission lines $x \to 0$ is best. For a model of the "ultimate" computer achieving thermodynamic limits on intercomponent information fluxes, the same mathematics as Chklovskii and Stevens's neuron model re-occurs for a different reason, so that again $x = 3/2$ seems best.**

Keywords — Neurons, fiber optics, superconducting transmission lines, computer architecture, VLSI theory, thermodynamic limits on information flux.

1 Neurons

Dmitri Chklovskii gave a talk at my institute on 14 April 1999 about "optimal wiring in the brain." One of the arguments in his talk concerned the relative amount of your brain's volume occupied by "wires," (axons) as opposed to "components" (neurons).

He argued as follows. The speed of signal propagation in non-myelinated (i.e. the kind predominant in your brain) axons is proportional to $D^{1/2}$ where D is the axon diameter. This proportionality comes from both a theoretical model and from experiments by Rushton in the 1950s. Assume your brain has an essentially fixed number of components of fixed size, a fixed number of wires, and a fixed wiring diagram. Assume that the packing efficiency is essentially invariant as the wire lengths and thicknesses change[1].

Given these constraints: optimize "performance," i.e. total bit rate, by choice of the ratio x of wire volume to component volume. If ℓ is the average wire length, then $\ell \propto (x+1)^{1/3}$ assuming the lengths scale as volume$^{1/3}$. If the wire diameters are D, then $D^2\ell \propto x$ by considering the total volume occupied by wires, hence $D \propto x^{1/2}(x+1)^{-1/6}$. Then the performance (assuming at most 1 bit is in a wire at any one time) is proportional to $D^{1/2}/\ell$, i.e. to $x^{1/4}(x+1)^{-5/12}$. Upon maximizing performance by choice of x, I find $x = 3/2$, i.e. the wires:components volume ratio should be 3:2. This result agrees with Chklovskii's.

Chklovskii informs me that the volume ratio for "wires" versus "components" one actually sees when examining human brain tissue really is about 3:2.

Note: it is irrelevant to our present and future arguments just what the values of the proportionality constants are; the same exact optimal volume ratio will result regardless.

I then asked myself: what is the analogous argument going to tell us when applied, not to your brain, but to VLSI ("very large scale integration") circuitry?

2 2D VLSI, thin sheet conductors

For 2D VLSI circuitry, if the "wires" are actually 2D patterns cut out of thin sheets of constant thickness, then the resistance of the

[1] This assumption is debatable. The packing efficiency of equal infinite cylinders is known to be $\pi/\sqrt{12} \approx .9069$, achieved by the vertices of the equilateral triangle lattice; and the packing efficiency of equal 3D balls is known to be $\pi/\sqrt{18} \approx .7405$, achieved by the FCC lattice. That suggests that packing efficiencies will be $> 70\%$, and at least that that any variation will not change anything by $\gtrsim 30\%$.

wires of length ℓ and width W is proportional to ℓ/W, whereas the capacitance of a wire is proportional to ℓW. (The capacitance of a 2-plate capacitor is $\epsilon_0 A/d$ where A is the area of each plate and d is the gap between the plates.) Hence the RC delay time is proportional to ℓ^2 independent of width. (Note the propagation time is *not* linear in ℓ; signal propagation in resistive-capacitive wires is diffusive rather than wavelike.) Also, the capacitance charge-up time from a bounded-current source, is proportional to ℓW. Both of these things tell us that it is best to make all the wires as thin as possible, i.e. to strive for $x = 0$. Actually in practice, unfortunately it is impossible to make the wires much thinner than the components (MOSFETs) and wiring unavoidably consumes a substantial fraction of chip area, i.e. x is considerably larger than 0. Indeed in the early 1980s it was feared that wires would take over as VLSI circuitry got more complicated, consuming almost all of the chip area. But as of 1999 those fears have not been realized, due to better routing and layout algorithms and to the appearance of VLSI processes with more wiring layers[2].

3 3D VLSI, CYLINDRICAL WIRES

For 3D VLSI, the situation is more interesting. The capacitance of a capacitor which is two concentric cylinders of radii a, b (this models a wire) and length ℓ long ($a < b \ll \ell$) is $2\pi\epsilon_0\ell/\ln(b/a)$. Meanwhile the resistance of an L-long wire of radius a is $\rho\ell a^{-2}\pi^{-2}$, where ρ is the resistivity of the metal (assumed constant). Hence making the wires thicker will reduce the RC delay substantially and only increase the C-charge time by a logarithmic factor (actually it won't increase it at all if b grows proportionally to a), suggesting it is a good idea to use wires of nonminimal thickness. More precisely, we have:

If ℓ is the average wire length, then $\ell \propto (x+1)^{1/3}$ assuming the lengths scale as volume$^{1/3}$. If the wire diameters are D, then $D^2\ell \propto x$, hence $D \propto x^{1/2}(x+1)^{-1/6}$. Then the performance is proportional to $D^2\ell^{-2}$ if the main factor responsible for delay is thought to be RC delays in wires. This is $\propto (x+1)^{-1}x$ which actually causes the optimum to be at $x = \infty$, i.e. you always do better by making the wires thicker! Of course at some thickness the approximate treatment of the wires as long thin cylindrical capacitors will lose relevance, so the best x won't really be infinite, but it will be large. But if instead the main delay-causing factor is the limited current-drive capability of the components, then the performance is proportional to ℓ^{-1}. In that case you are best off making the wires as thin as possible.

The optimal tradeoff between these two effects will happen approximately when the wires are made so thin that their resistance starts to become larger than the source resistance of their drivers. More precisely, supposing $b = 2a$, the RC delay in a wire of length ℓ and outer radius b is $(R_{\text{source}} + 4\rho\ell b^{-2}\pi^{-2})2\pi\epsilon_0\ell/\ln 2$. If $\ell = (x+1)^{1/3}K_1$ and $b = x^{1/2}(x+1)^{-1/6}K_2$ then the optimal x obeys

$$R_{\text{source}}^3 K_2^6 x^6 \pi^6 = 12^3 \rho^3 K_1^3 (x+1)^2. \qquad (1)$$

An interesting consequence of all this, which A. Gottlieb tells me was pointed out to him once by computer architect Burton Smith, is that if you have a cube of side ℓ and want to transmit information from one face of the cube to the other using parallel arrays of cylindrical resistive-capacitive wires, then the maximum possible bandwidth one can achieve is about $(2\epsilon_0\rho)^{-1}\ln 2$ bits per second *independent* of ℓ and the amount of power available. With $\rho = 1.587 \times 10^{-8}$ohm-meter (the resistivity of Silver at 293 Kelvin), this is

$$2.46 \times 10^{18} \text{ bits/second.} \qquad (2)$$

4 2D VLSI, CYLINDRICAL WIRES

If in 2D VLSI, we now view the wires as 3D cylinders, not 2D patterns cut out of thin sheets, then:

If ℓ is the average wire length, then $\ell \propto (x+1)^{1/2}$ assuming the lengths scale as area$^{1/2}$. If the wire diameters are D, then $D\ell \propto x$, hence $D \propto x(x+1)^{-1/2}$. Then the performance is proportional to $D^2\ell^{-2}$ if the main factor responsible for delay is thought to be RC delays in wires. This is $\propto (x+1)^{-2}x^2$ again causing the optimum to be at $x \to \infty$. But if

[2] Incidentally, one may ask: what about inductance of the wires, and what about the fact that at high frequencies, almost all of the current is carried near the surface of a wire ("skin effect") causing the effective resistance to be considerably higher than I'd said it was? The answer to both of these questions is that such inductive effects are usually neglectable for long and/or thin wires; the scaling laws are such that RC effects dominate. The inductance of a long cylindrical wire of radius a (supposing all the magnetic flux encircling it is contained in a concentric coaxial tube of radius b, $b > a$) is $(2\pi)^{-1}\mu_0 \ln(b/a)$ per unit length. If b/a is held fixed, then for a long thin wire, the typical $\sqrt{LC}$ delay, where L denotes the inductance and C the capacitance of the wire, is ℓ/c, where ℓ is the length of the wire and c is the speed of light. Meanwhile, the RC delay (where R is the resistance) is typically far larger and grows proportionally to $\ell^2 a^{-2}$, as opposed to $\ell^1 a^0$, as ℓ increases and a decreases.

instead the main delay-causing factor is the limited current-drive capability of the components, then the performance is proportional to ℓ^{-1}. In that case you are best off making the wires as thin as possible: $x \to 0$.

Again the optimal tradeoff between these two effects will happen approximately when the wires are made so thin that their resistance starts to become larger than the source resistance of their drivers. More precisely,

More precisely, supposing $b = 2a$, the RC delay in a wire of length ℓ and outer radius b is $(R_{\text{source}} + 4\rho\ell b^{-2}\pi^{-2})2\pi\epsilon_0\ell/\ln 2$. If $\ell = (x+1)^{1/2}K_1$ and $b = x(x+1)^{-1/2}K_2$ then the optimum x obeys

$$256\rho^2 K_1^2(x+1) = R_{\text{source}}^2 \pi^4 x^6. \tag{3}$$

5 Fiber optics and superconducting transmission lines

Finally for 3D (or 2D) VLSI using fiber optics rather than wires to carry signals, the signal speed is $c\eta$ (where η is the constant refractive index of glass, and c is the constant speed of light) so the best thing to do again would be just the obvious course of making the fibers as thin as possible (namely, of order 1 wavelength of light) but of course no thinner (which would prevent the fibers from working). This same conclusion also would pertain to "superconducting transmission lines" instead of wires or fibers, and these conclusions are not affected if we allow more than one bit to co-reside in the same fiber at one time, so long as making the fibers thicker is regarded as not affecting the number of bits per unit length of fiber. (Allowing more than 1 bit is different from our previous models of wires and neurons, where a wire could contain at most 1 bit at any time.)

6 Ultimate thermodynamic limits on information flux

In a previous work [3] I showed that the maximum sustainable bit rate $\dot{I}$ transmissible through an area-A window with power P, obeyed the law $\dot{I}/A < (P/A)^{3/4} \cdot \kappa$, at least in the high flux limit. Here $\kappa \approx 2.149 \times 10^{21}$ sec$^{5/4}$Kg$^{-3/4}$meter^{-2}bit is a fundamental physical constant. This is by full use of all the distinguishable quantum states of a radiation field. (Essentially, the bound becomes tight when the radiation has the same entropy as "blackbody radiation" at the appropriate "temperature" to achieve that power level. If it were possible to transmit information at a higher rate than the bound, then I showed one could obtain a contradiction with the second law of thermodynamics.) In other words, if one has a fixed amount of power and wants to transmit bits through a "pipe" of diameter D, the maximum achievable sustainable bit rate is proportional to $D^{1/2}$.

In this scenario (unlike in the models of neurons and wires) we allow many bits to co-reside in the same pipe at the same time. In that case, if we wish to maximize total bit flow rate and do not care about propagation times ("latencies"), then the lengths of the pipes are not directly relevant, in the sense that a twice as long pipe will still have the same bit flow rate. Hence, we might as well make the pipe diameters D as large as we can (at least, up to the point where the power flux per unit area decreases to the point where my law is no longer valid).

On the other hand, suppose we care about both bandwidth and latency. In particular, suppose we want to maximize the bit rate divided by the latency – which for one pipe is proportional to $D^{1/2}/\ell$. In this case we will have the same mathematical problem as Chklovskii's for non-myelinated neurons (although for completely different reasons)! Therefore, the same conclusion holds as for Chklovskii; the total volume of the pipes is best chosen to be 3/2 times the total volume of the components.

Incidentally, if we are really arguing here about the ultimate limits set by the laws of physics on computers, then one may quibble: why must ultimate computers be made of information "pipes?" Perhaps they use free-space information transmission. However, if energy density and component density are both bounded and bitrate from each component is bounded, then an argument of Ozatkas and Goodman [2] suggests but does not prove[3] that the computer *must* consume at least the same amount of volume as would be occupied by *pipes* of constant cross sectional area to

[3] Unfortunately, although Ozatkas and Goodman prove a nice statement, their subsequent claim that "therefore... all the results of VLSI area theory apply also to optics" is overconfident. O&G's "ℓ_{tot}" refers to the total path length actually followed by the light. Wires cannot necessarily follow the paths followed by light (i.e. if light self-intersects).

For instance in 2D VLSI, laying out the complete graph K_N takes area $\theta(N^4)$, whereas O&G's 2D bound, if all the N nodes are located in a square subregion of area N, gives order $N^{2.5}$. Hence, optics, even at a single wavelength, could still be better than wires, as far as O&G's bound is concerned, despite their claims.

(Maybe this counterexample can be defeated, and maybe O&G's statement is true, but I'm just saying: they never proved it.)

transmit the same amount of information. Essentially, at the optimum, the whole of space is filled with information-carrying light waves at a fixed energy density.

7 What function to maximize?

Of course in §6 , it is debatable whether the function one wants to maximize to get a good general purpose computer, really is the bit rate B divided by the latency T. I asked this question (of what function $F(B,T)$ to maximize) of some computer architects and it was difficult to get an answer! They all said things like "for different algorithms, latency or bandwidth may matter more; one has to select a reasonable tradeoff based on the mix of algorithms one expects to be running" – true enough, but what about demanding an end to such weaseling and demanding those architecture experts provide a *specific* function of bandwidth and latency (which they feel encapsulates such tradeoffs) to maximize? I prefer B/T, basically because it is the simplest choice. Harold Stone suggested (essentially) B^p/T where p is a very large constant (i.e. he cares very much about B and very little about T). Both choices have been examined in §6.

In conclusion, let me note that I have restricted myself to maximizing information-throughput "performance" measures subject to volume constraints and have completely *ignored* other considerations, such as heat production and consequent cooling requirements.

References

[1] M. Chklovskii & C. Stevens: Manuscript, submitted to Nature Neuroscience 1999.

[2] H.M. Ozatkas & J.W. Goodman: Lower bound for the communication volume required for an optically interconnected array of points, J. Optical Soc. America 7,11 (Nov. 1990) 20-26.

[3] Warren D. Smith: Fundamental physical limits on computation NECI tech. report 1992.