JOURNAL OF THE ROYAL SOCIETY Interface

Scaling theory for information networks

Melanie E Moses, Stephanie Forrest, Alan L Davis, Mike A Lodder and James H Brown

J. R. Soc. Interface 2008 **5**, 1469-1480
doi: 10.1098/rsif.2008.0091

JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2008) **5**, 1469–1480
doi:10.1098/rsif.2008.0091
Published online 9 May 2008

Scaling theory for information networks

Melanie E. Moses[1,*], Stephanie Forrest[1,4], Alan L. Davis[2], Mike A. Lodder[2] and James H. Brown[3,4]

[1]*Department of Computer Science, and* [3]*Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA*
[2]*School of Computing, University of Utah, Salt Lake City, UT 84112, USA*
[4]*Santa Fe Institute, Santa Fe, NM 87501, USA*

Networks distribute energy, materials and information to the components of a variety of natural and human-engineered systems, including organisms, brains, the Internet and microprocessors. Distribution networks enable the integrated and coordinated functioning of these systems, and they also constrain their design. The similar hierarchical branching networks observed in organisms and microprocessors are striking, given that the structure of organisms has evolved via natural selection, while microprocessors are designed by engineers. Metabolic scaling theory (MST) shows that the rate at which networks deliver energy to an organism is proportional to its mass raised to the 3/4 power. We show that computational systems are also characterized by nonlinear network scaling and use MST principles to characterize how information networks scale, focusing on how MST predicts properties of clock distribution networks in microprocessors. The MST equations are modified to account for variation in the size and density of transistors and terminal wires in microprocessors. Based on the scaling of the clock distribution network, we predict a set of trade-offs and performance properties that scale with chip size and the number of transistors. However, there are systematic deviations between power requirements on microprocessors and predictions derived directly from MST. These deviations are addressed by augmenting the model to account for decentralized flow in some microprocessor networks (e.g. in logic networks). More generally, we hypothesize a set of constraints between the size, power and performance of networked information systems including transistors on chips, hosts on the Internet and neurons in the brain.

Keywords: metabolic scaling; networks; microprocessors

1. INTRODUCTION

Computer networks span a vast range of physical sizes and can have billions of components. For example, modern microprocessors contain billions of transistors networked in a few square centimetres of surface area (C. Ludloff 2007, http://sandpile.org/). The Internet connects half a billion hosts (Internet Systems Consortium 2008, http://www.isc.org/index.pl?/ops/ds/host-count-history.php) and spans the $5\times10^7\,\mathrm{km}^2$ surface of the Earth. The size of a network can be measured in terms of the number of nodes, in which case the Internet is about the same size as a computer chip, or as the physical distance spanned by the links, in which case the Internet is 10 orders of magnitude larger.

Although computational complexity theory describes the scaling properties of algorithms, and very-large-scale integration (VLSI) scaling describes how some properties of computer chips vary with process and transistor size (Mead & Conway 1979), computer science lacks a general predictive theory of network scaling. For example, how does wire length or power consumption scale with the number of transistors on a chip? How will bandwidth demand and latency change as more hosts are added to the Internet? Minimizing power consumption and predicting bandwidth demand are important questions in the design of these systems. Answers to such questions require a theory of network scaling: one that describes how power consumption, latency and the physical footprint of a network scale as functions of system size, the number of components and the degree of centralization.

Organisms also use networks to distribute energy, materials and information to individual components. We adapt the principles of metabolic scaling theory (MST; West *et al.* 1997; Brown *et al.* 2004) in order to describe the scaling of engineered information networks. The term 'MST' is used throughout the paper to refer to the original West, Brown and Enquist model (West *et al.* 1997). MST is extended to accommodate variation in the size and density of network components. We also discuss how decentralized network designs deviate from MST predictions that assume a single central source of flow in the network. 'The last

*Author for correspondence (melaniem@cs.unm.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2008.0091 or via http://journals.royalsociety.org.

Received 2 March 2008
Accepted 18 April 2008

Table 1. Glossary of notation.

symbol	meaning
A	cross-sectional area
vol	volume
M	mass
P	power (active power)
V	voltage
B	metabolic rate
D	dimension
t	time
b	scaling exponent
N_x	number of components of type x
L	length of an edge in a network
K	depth of a tree
k	subscript denoting a level in a hierarchical branching network ($k=0$ refers to the root)
β	branching ratio N_{k+1}/N_k in a network (also called fanout)
γ	length ratio L_k/L_{k+1}
w_k	width of an edge at level k in the network (e.g. wire width)
ω	width ratio w_k/w_{k+1}
ρ	density (per unit area or volume)
λ	process width
net	subscript denoting an entire network
org	subscript denoting an entire organism
chip	subscript denoting an entire computer chip
c	subscript denoting the lowest level component in a branching structure (e.g. capillary)
tr	subscript denoting transistors

mile' (the interface between the network and the components) and the degree of centralization in networks emerge as important topics for future research in both biological and information systems.

The paper is organized as follows. We review MST and explain its implications for information networks, using the particular example of networks known as H-trees that are often used to distribute clock signals on microprocessors. We then show how certain engineering trade-offs cause systematic differences between network scaling in microprocessors and organisms. We conclude with a discussion of how the MST approach offers a valuable theoretical perspective to guide and assess the design of many networked information systems.

1.1. Network scaling in organisms

Scaling describes how some property of a system, the dimension of a structure or the rate of a process, varies systematically with some other property. Many scaling relations are described by power functions of the form $Y = Y_0 X^b$, where Y is a dependent variable, Y_0 is a normalization constant, X is the independent variable and b is the scaling exponent. Our notation is summarized in table 1. We often ignore the normalization constant and write $Y \propto X^b$. The scaling exponent quantifies how one component of a system changes with respect to another, e.g. how the volume of the circulatory system grows to service an increasing number of cells, or how the footprint of wire in a computer chip must grow to service an increasing number of transistors. In biology, b is often a simple multiple of 1/4; for example, metabolic rate (B) scales with body mass (M) as $B \propto M^{3/4}$; blood circulation times, gestation period and lifespan (t) scale as $t \propto M^{1/4}$ (Kleiber 1932; Peters 1983; Schmidt-Nielsen 1984; West *et al.* 1997).

MST posits that quarter-power exponents arise from optimized hierarchical branching resource distribution networks, such as mammal and plant vascular systems. These designs evolved under natural selection because they minimize the energetic cost of transporting resources while balancing the competing concerns of performance, speed and efficiency. According to MST, the delivery capacity of vascular networks scales with the 3/4 power of the volume of the network. Thus, larger networks in larger organisms deliver more total energy, but with diminishing returns. Because larger networks deliver less energy proportional to their size, rates of cellular metabolism are forced to decrease with increasing body size. An alternative design in which the cellular metabolic rate remained invariant and the whole organism's metabolic rate increased linearly with body size would require the volume of the network to increase nonlinearly as the 4/3 power of body mass. It is easy to see why natural selection has not adopted this strategy. If a 2 g shrew has a reasonable volume of blood to support its metabolism, say 0.1 g, a 2 ton elephant will require 100 tons of blood, an obvious impossibility.

The quarter-power scaling relation is derived mathematically from the laws of physics (e.g. conservation of matter and energy and the laws of hydrodynamics) and four fundamental constraints on the centralized, hierarchical networks that supply raw materials (oxygen and energy) for metabolism (West *et al.* 1997), restated as follows:

— *A. Cross-sectional area preserving.* The summed cross-sectional area (A_{net}) at each hierarchical level is constant throughout the network, so that flow rate is constant and impedance is minimized. Thus, $A_{\mathrm{net}} = A_k N_k$ for all k where A_k is the cross-sectional

area of each branch at level k, and N_k is the number of branches at level k. This implies that for the final branch of the network (the capillaries) $A_{\rm net}=A_{\rm c}N_{\rm c}$. We note, however, that area preservation is violated in the lowest levels of the hierarchy of the cardiovascular system to slow blood velocity through the capillaries and allow for diffusion of oxygen (West *et al.* 1997).

— *B. Space-filling.* A hierarchical, self-similar and space-filling network supplies resources to the entire three-dimensional volume of the body. One way to achieve this is with a self-similar (or fractal) rule in which the lengths of branches decrease systematically lower in the hierarchy. MST defines a length ratio as

$$\gamma=\frac{L_k}{L_{k+1}}=\beta^{1/3}, \tag{1.1}$$

where L_k is the length of an edge at level k in the hierarchy, and β is the number of daughter branches per parent branch. According to MST the 1/3 exponent arises from self-similarity in three dimensions, which requires each branch to be proportional to the radius of the volume of tissue to which it delivers energy (West *et al.* 1997). In the electronic supplementary material, we derive the MST prediction (for the two-dimensional case) for the length, $L_{\rm net}$, from the centre of the network to any leaf (i.e. the sum of all edge lengths from the aorta to any capillary, assuming a balanced tree). In three dimensions the prediction is

$$L_{\rm net}\propto N_{\rm c}^{1/3}L_{\rm c}, \tag{1.2}$$

where $L_{\rm c}$ is the length of a capillary and $N_{\rm c}$ is the number of capillaries. Since $N_{\rm c}$ is the number of leaves in the tree, then $N_{\rm c}=\beta^K$, where K is the depth of the tree.

— *C. Terminal units are invariant.* The lengths ($L_{\rm c}$), cross-sectional areas ($A_{\rm c}$) and delivery capacity of capillaries do not vary systematically with M. Assuming both the length ratio of equation (1.1) and invariant terminal units is a controversial aspect of MST (Makarieva *et al.* 2005*a*; Etienne *et al.* 2006), and we modify it in our analysis of microprocessors in §3.1.

— *D. Network volume (vol_{net}) is proportional to M.* Optimization principles (West *et al.* 1997) and empirical evidence (Peters 1983) show that, for example, blood volume is approximately 7% of organism volume in mammals of all sizes. Because organism volume is proportional to M, then $\mathrm{vol}_{\rm net}\propto M$. This constraint is also modified to cover the case of microprocessors.

From the four constraints above, MST predicts how $\mathrm{vol}_{\rm net}$ scales with $N_{\rm c}$, the number of leaves of the tree. Given a balanced tree with cross-sectional area preserving, $\mathrm{vol}_{\rm net}=L_{\rm net}A_{\rm net}$. From constraints (i) and (iii), and equation (1.2), $\mathrm{vol}_{\rm net}\propto A_{\rm c}L_{\rm c}N_{\rm c}^{4/3}$. More generally, in dimension D,

$$\mathrm{vol}_{\rm net}\propto A_{\rm c}L_{\rm c}N_{\rm c}^{(D+1)/D}. \tag{1.3}$$

Because the constraints require $\mathrm{vol}_{\rm net}\propto M$, and $A_{\rm c}$ and $L_{\rm c}$ to be constant, $N_{\rm c}^{(D+1)/D}\propto M$, allowing us to predict the number ($N_{\rm c}$) and density ($\rho_{\rm c}$) of capillaries in an organism as follows: $N_{\rm c}\propto M^{D/(D+1)}$ and $\rho_{\rm c}\propto M^{-1/(D+1)}$. Since each capillary delivers energy at the same rate, this gives the canonical 3/4 power scaling equation for metabolic rate, $B\propto N_{\rm c}\propto M^{D/(D+1)}$ where $D=3$.

Through these four constraints, MST provides a mathematical explanation for $D/(D+1)$ scaling in biology. However, there are several controversies surrounding MST. These fall into two general categories. First, several authors suggest that the 3/4 power scaling of metabolism is not universal, citing examples from particular taxonomic groups, frequently mammals (White & Seymour 2003), and also across animals more generally (Glazier 2005; Makarieva *et al.* 2005*b*; but see Savage *et al.* (2004) and Moses *et al.* (2008)). Similar controversy surrounds the assertion of a universal relationship between temperature and metabolic rate. Regardless of whether there is universal network scaling in biology or whether MST is accepted as the best explanation for it, the scaling approach may still be appropriate in the domain of microprocessors and other information networks.

More relevant to this work is a second set of criticisms that the model assumptions are either violated, internally inconsistent or unnecessary to generate $D/(D+1)$ scaling. In particular, several authors have questioned the assumption of invariant terminal units, suggesting that capillaries may vary in size or, more importantly, may vary systematically with body size (Dawson 2003; Kozlowski & Konarzewski 2004; Makarieva *et al.* 2005*a*). That criticism is addressed in §3.1.1, where we modify the theory to incorporate the widely varying sizes of transistors into our model. We further relax the assumption that the length of the terminal wire in a network is constant. In our model, we estimate the distance between transistors directly from their density. Some have questioned whether self-similar space filling (constraint B) is theoretically justified (Etienne *et al.* 2006) and whether it is consistent with the assumption of invariant terminal units (Kozlowski & Konarzewski 2004), but see Brown *et al.* (2005). In §3.1 we show that the hierarchical H-tree that distributes clock signal on a chip follows the self-similar space-filling predictions when we allow terminal units to vary in length. Finally, alternative models suggest that 1/4 powers can be generated for any resource distribution network without requiring fractal branching (Banavar *et al.* 1999). We adopt the fractal branching model in this paper because it describes the H-tree so well. However, more general models that do not require fractal branching may be appropriate to describe other information networks. Despite these controversies, we believe that the MST formalism is relevant to computing and provides a starting framework for articulating and testing hypotheses about how such engineered systems scale.

1.2. Wire scaling on computer chips

Since microprocessors were first built in the early 1970s, the number of transistors on an integrated circuit has increased by five orders of magnitude (Mezhiba & Friedman 2002; Moore 2003). The increase is due

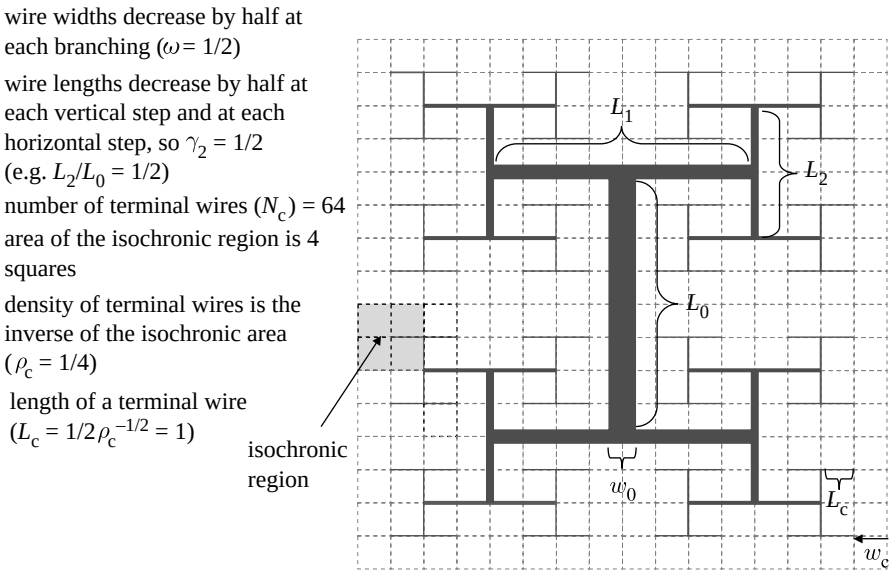

Figure 1. The H-tree, a two-dimensional hierarchical self-similar branching network that distributes clock signals on an integrated circuit.

largely to shrinking the size of transistors (measured in process width, λ) by over 100-fold and increasing chip area (A_{chip}) by approximately 30-fold (see electronic supplementary material). The increase in transistor density (ρ_{tr}), and commensurate increase in information processing power (measured in millions of instructions per second, MIPS), has required a 500-fold increase in the power to operate each chip, from approximately 0.2 W in 1978 to over 100 W today in Intel microprocessors (Moore 2003). Consequently, power and heat dissipation are now dominant constraints on chip design. The total area consumed by wires (A_{net}) has increased superlinearly as a fraction of A_{chip}, and this is a significant component of the growth in power requirements. We examine how well MST principles characterize the relationship between A_{chip}, λ, transistor density (ρ_{tr}) and power (P).

Equations (1.1)–(1.3) can be used to derive the scaling behaviour of networks that distribute clock signals on computer chips. We treat wires that deliver charge to transistors analogously to the arteries that deliver oxygen to cells. The wire network is subject to physical laws (e.g. conservation of current as described by Kirchoff's Laws). Important properties of the chip, such as delay, heat generation and power consumption, are influenced by the physical lengths and cross-sectional areas of the wires. The flow of current through the network of wires and transistors is analogous to the flow of blood through the circulatory system.

Despite these similarities, there are important differences between circulatory networks and wires on chips. First, organisms are three dimensional, but chips are often referred to as 2.5 dimensional (Deng & Maly 2004) because the area covered by transistors is two dimensional, but metal layers of wires extend into a relatively thin third dimension. Second, the circulatory system is fully centralized, with all blood originating from and returning to the heart, but the logic networks on microprocessors are somewhat decentralized in that not all signals pass through any one part of the network.

To simplify the translation of MST to chips, we focus on the H-tree (shown in figure 1), which is a fully centralized two-dimensional clock distribution network. The mapping between the vertebrate circulatory system and H-trees is straightforward. The simplest H-tree design has long, wide wires that branch into successively shorter and narrower wires in quantitative agreement with MST predictions for a two-dimensional network. Both are hierarchical, fractal-like, branching trees that deliver resource (clock signal or blood) from a central source (clock or heart) to the leaves of the tree (terminal clock buffers or capillaries) that are distributed throughout the area (or volume) of the system. The terminal clock buffers (or capillaries) deliver signal (or oxygen) to an area of chip or volume of tissue. The area to which a single clock buffer delivers signal is called the isochronic region and is important in our theory. Finally, we refer to the region where the terminal units deliver signals or energy to the components of the system (transistors or mitochondria) as *the last mile* by analogy with telecommunications networks. Both networks enable the system to function as an integrated whole by ensuring that resources are distributed to all regions of the system.

The clock distribution system ideally presents a clock signal to all points of the chip simultaneously. In practice, simultaneity is impossible due to small variations in process, voltage and temperature (PVT) at different locations on the chip. The maximum difference in clock arrival times under worst case conditions is called skew, and H-trees illustrate a trade-off between minimizing skew and minimizing power (Friedman 2001). Although other clock designs require less power, the H-tree provides identical path lengths from the clock source to the leaves of the tree, reducing skew. Small differences in path lengths occur, primarily due to slight geometric irregularities induced by cell set variations and automated design tool algorithms. This is important for high-frequency operation, because skew is combined with the effective

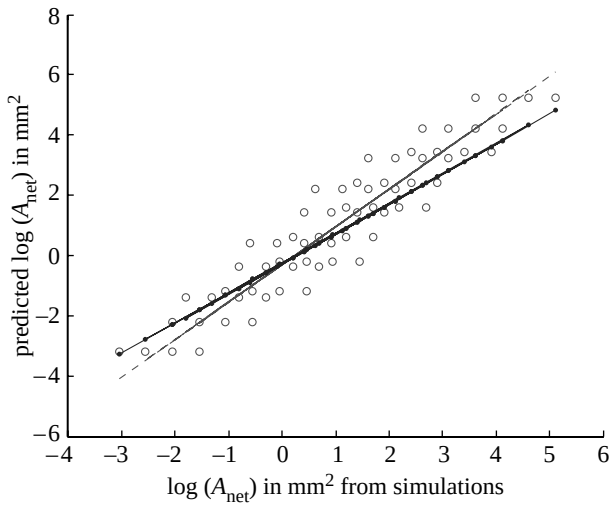

Figure 2. Predicted versus simulated A_{net}. The open circles show the predictions of equation (3.2), which assumes invariant terminal units. The filled circles show the predictions of equation (3.3), which sets the lengths of terminal wires such that they reach the centre of the isochronic region. The slope close to 1 and high r^2 indicate that equation (3.3) is a more accurate scaling prediction. Dashed line, RMA slope = 1.25, $r^2 = 0.88$; solid line, RMA slope = 0.99, $r^2 = 0.999$.

delay due to PVT variations and to worst case logic and wire propagation delays to determine the shortest clock period (or the highest frequency) at which the chip will operate correctly. The fractal branching network described by MST actually describes idealized H-trees better than evolved biological networks because the balanced tree design eliminates the possibility of asymmetric side branches in the network.

2. METHODS

We obtained measurements of active power (P), clock frequency (f), process size (λ), chip area (A_{chip}) and number of transistors (N_{tr}) for 516 microprocessors manufactured by Intel, AMD, Via and Cyrix from 1991 to 2006 (see electronic supplementary material). We included only those chips for which a single value of every variable was given, and we excluded multi-core chips. Most chips have several releases with different specifications, and each of these appears separately in our dataset. We also included data for each of seven chips manufactured by Intel in the 1970s and 1980s (the 4004 through the 486; Moore 2003; electronic supplementary material). In the case of the H-tree, we do not have direct measurements of the total wire area (A_{net}) or density of the terminal units. Thus, the data reported in figure 2 were generated by simulations described below.

We follow standard methods to determine the scaling exponents from data by taking the log (base 10) of both variables and calculating a linear regression through the log-transformed data. Different regression techniques can lead to different scaling exponents (Warton *et al.* 2006; O'Connor *et al.* 2007). Although ordinary least-squares (OLS) has been traditionally used, it assumes that all errors occur in the dependent variable. Reduced major axis (RMA) regression assumes that error is equally distributed between the dependent and independent variables. Since we do not know how error is distributed in these data, we report exponents obtained from both RMA and OLS regression with 95% CIs for each. We suggest that these estimates provide reasonable upper and lower bounds for the true scaling exponent.

3. RESULTS

3.1. Scaling of the H-tree

In this section we translate constraints A (area preserving) and B (space filling) to two dimensions to predict scaling of a simple two-dimensional H-tree. Then we show that clock trees violate constraints C and D and how that alters scaling properties.

— *A. Wire width.* Cross-sectional area preserving Constraint A becomes wire width preserving for the two-dimensional clock tree when wires are kept on two metal layers that have constant wire thickness. The summed width of wires at each hierarchical level allows impedance matching (minimizing signal reflections) and maintains constant current density with minimal resistance throughout the network.
Prediction. We define the wire width ratio, $\omega = w_k/w_{k+1}$, and note that $\omega = \beta$ owing to the area preserving constraint, where w_k is the width of each wire at level k and β equals fanout. For an H-tree, ω and fanout are equal to 2 at all nodes. Thus, the model predicts that the wire width at each level will be 1/2 the wire width at the level above.
Observation. Designs for efficient H-trees (Friedman 2001) specify $\omega = 2$ (figure 1). These specifications require that the sum of wire widths (w_{net}) at each hierarchical level be constant, and $w_{net} = w_c N_c$. As an example in figure 1, $w_c = 1$ and $N_c = 64$, so that $w_{net} = 64$. The same value of w_{net} is obtained for $k = 0$ since $w_0 = 32$ and $N_0 = 2$.

— *B. Wire length.* The biological constraint B (networks are space filling) translates to area filling for two-dimensional chips. Each parent H branches twice to create a new child H, so we consider the scaling of parallel branches of the H, each of which is two branching generations from its parent (for example, L_2 and L_0 in figure 1).
Prediction. In two dimensions, accounting for the two branchings to generate each H, we rewrite equation (1.1) as

$$\gamma_2 = L_k/L_{k+2} = (2\beta)^{1/2}. \tag{3.1}$$

Given $\beta = 2$, then $\gamma_2 = 2$; the length of the four wires in a child H will be half the length of wires in the parent H.
Observations. The length of wires in each daughter H is 1/2 that of the parent H (figure 1), following MST length scaling predictions.

— *C. Network area.* For a perfectly balanced tree with area-preserving branching, $A_{net} = w_{net} L_{net}$. Using the length ratio in equation (3.1) and the

electronic supplementary material, we obtain a two-dimensional version of equation (1.2) for network length (distance from the clock to any one leaf): $L_{\mathrm{net}} \propto N_{\mathrm{c}}^{1/2} L_{\mathrm{c}}$. Thus, the two-dimensional version of equation (1.2) is

$$A_{\mathrm{net}} \propto w_{\mathrm{c}} L_{\mathrm{c}} N_{\mathrm{c}}^{3/2}. \tag{3.2}$$

Equation (3.2) shows that if the lengths (L_{c}) and widths (w_{c}) of terminal wires are constant, the area of chip allocated to wire grows faster than the area allocated to components.

Accounting for variation in the size of terminal units. In MST A_{c} and L_{c} are invariant across body sizes, but in chips w_{c} and L_{c} may vary substantially. The minimum w_{c} is determined by process size (λ). We assume $w_{\mathrm{c}} \approx \lambda$, although in some cases w_{c} is slightly greater than λ due to artefacts of the design process and slightly irregular geometries created by variations in logic complexity; this has negligible effect on the predictions. The lengths of the terminal H-tree wires (L_{c}) also vary. When the density of terminal wires (ρ_{c}) of the H-tree is lower, the terminal nodes are further apart, so the wires must be correspondingly longer in order to create a single connected H-tree. Dimensional analysis gives $L_{\mathrm{c}} = 1/2\rho_{\mathrm{c}}^{-1/2}$, where, by definition, $\rho_{\mathrm{c}} = N_{\mathrm{c}}/A_{\mathrm{chip}}$. This simply means that the terminal clock wire is half the length of the isochronic region, and therefore reaches the centre of the isochronic region as shown in figure 1. Additionally, the wire at the next hierarchical level is twice as long and is therefore able to connect terminal wires in adjacent isochronic regions.

Prediction. Substituting in the variables w_{c} and L_{c}, the network area equation becomes

$$A_{\mathrm{net}} \propto \lambda N_{\mathrm{c}} A_{\mathrm{chip}}^{1/2}. \tag{3.3}$$

If component density (ρ_{c}) is held constant, then equations (3.2) and (3.3) are equivalent. However, if ρ_{c} varies across chips, then equation (3.3) more accurately describes the scaling of H-trees.

Observations. We do not have access to direct measurements of A_{net} for clock trees on commercial chips. However we can simulate clock trees based on the design shown in figure 1. We simulated 64 H-trees by varying $\lambda = [10, 100, 1000, 10\,000]$ µm, $N_{\mathrm{c}} = [16, 16, 256, 4096, 65536]$ isochronic regions and $A_{\mathrm{chip}} = [10, 100, 1000, 10\,000]$ mm^2. We measured A_{net} by summing the area of all wires in the simulation for all combinations of these variables. We compared the measurements of A_{net} with the predictions of equations (3.2) and (3.3). Figure 2 shows that equation (3.3) is linearly related to the simulations with a very high r^2, while equation (3.2) (which assumes invariant terminal units) has lower r^2 and a slope greater than 1, indicating systematic deviations from the data.

It would be beneficial to obtain these measurements for H-trees on manufactured microprocessors, particularly since the use of repeaters and restrictions on wire widths in different metal layers may alter the dimensions of real H-trees from the idealized design. We are pursuing such measurements for future analyses.

— *D. Component density.* In MST constraint *D* specifies $\mathrm{vol}_{\mathrm{net}} \propto \mathrm{vol}_{\mathrm{org}}$, which requires that the number of capillaries (N_{c}) scales as $M^{D/(D+1)}$, and the density of capillaries (ρ_{c}) scales as $M^{-1/(D+1)}$. Constraint *D* is likely to be violated in microprocessors because additional metal layers are available to hold wire area that scales superlinearly with the area of the chip.
Predictions. A direct translation of Constraint *D* to a two-dimensional network predicts $\mathrm{A}_{\mathrm{net}} \propto \mathrm{A}_{\mathrm{chip}}$, and $\rho_{\mathrm{c}} \propto \mathrm{A}_{\mathrm{chip}}^{-1/2}$. However, if constraint D is violated, then there should be no systematic relationship between ρ_{c} and A_{chip}. We do not have direct measures of ρ_{c} on different microprocessors. However, ρ_{c} is the inverse of the area of the isochronic region, so that frequency (f) should be proportional to the radius of the isochronic region. Thus, if we assume $\rho_{\mathrm{c}} \propto f^2$, then we expect no relationship between f and A_{chip}, whereas MST would predict $f \propto A_{\mathrm{chip}}^{-1/4}$.
Observation. To test these predictions, we regressed f against A_{chip} on our dataset of 523 microprocessors. The regression is only marginally significant ($p = 0.013$) and explains only 1% of the variation in the data ($r^2 = 0.01$). Thus, as we expected there is no meaningful scaling relationship between f and A_{chip}.

3.2. Scaling on microprocessors

In order to test the predictions of MST on real data, we extrapolate the H-tree predictions to account for more general properties of microprocessors for which data are available.

3.2.1. Scaling of transistor density. The analysis in the previous section suggests that there is little relationship between the density of clock registers and A_{chip}. We generalize this analysis by treating all transistors as the terminal units of the logic network on a microprocessor. Then, we can test whether there is a relationship between the density of transistors (ρ_{tr}) and A_{chip} by regressing the log of ρ_{tr} against the log of A_{chip}. As expected, the regression is not significant ($p = 0.9$). However, ρ_{tr} is correlated with the process size λ: $\rho_{\mathrm{tr}} \propto \lambda^{-2.2}$. Using OLS regression $r^2 = 0.92$ and the 95% CI is -2.2 to -2.3 (the RMA exponent is -2.3). This empirical scaling relationship is close to what would be expected if the footprint of transistors were a constant fraction of chip area. This is because a lower bound on total transistor footprint can be estimated by assuming that the area of each transistor is the square of the minimum feature size (when in fact some transistors may be larger); then the summed area of all transistors (A_{tr}) is $A_{\mathrm{tr}} \propto N_{\mathrm{tr}}\lambda^2$. To test whether A_{tr} (using this estimate) is a constant fraction of chip area, we regress the log of $A_{\mathrm{tr}}/A_{\mathrm{chip}}$ against the log of A_{chip}. The regression is marginally significant ($p = 0.03$ and $r^2 < 0.01$), which means that there is almost no correlation between the proportion of chip area occupied by transistors and A_{chip}, and the proportion of chip area occupied by transistors is constant with respect to A_{chip}. To recap, MST posits that component

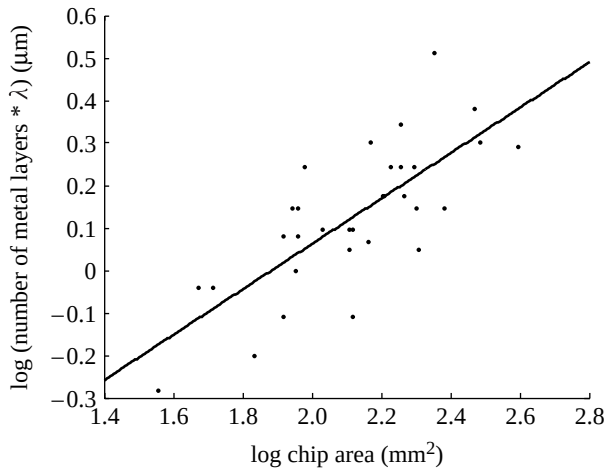

Figure 3. The number of metal layers increases with chip area as predicted by equation (3.3). The y-axis represents the number of layers containing wire (the number of metal layers plus the silicon layer) multiplied by the process size, and the x-axis represents chip area. The scaling exponent is 0.53 (OLS) and r^2 is 0.57. Data are from (C. Ludloff 2007, http://sandpile.org/).

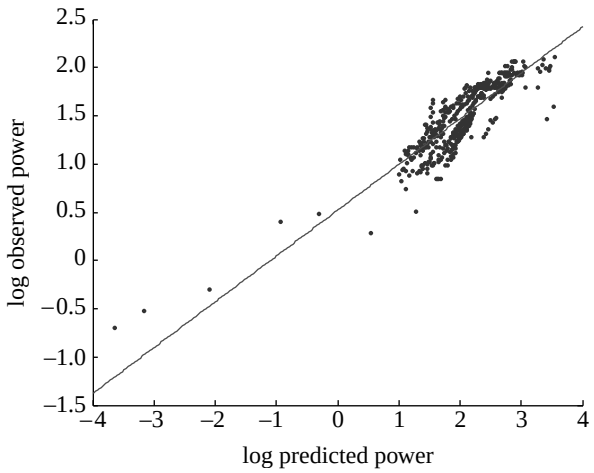

Figure 4. Power measured on microprocessors versus power predicted by equation (3.4) ($P \propto \lambda V^2 N_{\mathrm{tr}} A_{\mathrm{c}}^{1/2} f$). The scaling exponent is approximately 1/2. The exponent is significantly less than 1, indicating systematic differences between the scaling prediction and observations. The seven Intel chips from the 1970s and 1980s do not significantly affect the scaling exponent.

density decreases as the mass of the organism increases, but the data suggest that in microprocessors the density is unchanged as the chip area increases.

The 2.5-dimensional geometry of microprocessors may explain how transistor density is held constant as A_{chip} increases. Equation (3.3) can be used to predict how A_{net} increases with A_{chip} and therefore, how the number of metal layers increases with A_{chip}. Substituting $\rho_{\mathrm{tr}} \propto 1/\lambda^2$ into equation (3.3) gives $A_{\mathrm{net}} \propto A_{\mathrm{chip}}^{3/2}/\lambda$. Since the network of wires is contained on metal layers and on the silicon layer that holds transistors, we estimate A_{net} as $A_{\mathrm{net}} \propto (N_{\mathrm{metal}} + 1) A_{\mathrm{chip}}$, where N_{metal} is the number of metal layers. Thus, we predict $\lambda(N_{\mathrm{metal}} + 1) \propto A_{\mathrm{chip}}^{1/2}$. Figure 3 shows this relationship empirically. The OLS estimate of the scaling exponent is 0.53 (which is indistinguishable from the prediction of 1/2), and the RMA exponent is slightly higher (0.7). The data on metal layers (C. Ludloff 2007, http://sandpile.org/) include 32 microprocessors from the dataset described previously. These results show that, once λ is accounted for, the number of metal layers increases predictably with chip area. This correlation helps explain the reason for transistor density being not directly constrained by chip area: as area increases, the number of metal layers increases to accommodate excess wire area, reducing interference between routing wires and placing transistors.

3.2.2. Power scaling prediction. It is important to understand wire scaling because the footprint of the wire influences the amount of power required to operate a chip. Wire length is often estimated using Rents Rule (Stroobandt 2001), but we use scaling theory and estimate A_{net} from equation (3.3). We use this estimate of A_{net} to predict active microprocessor power, and then we test those predictions against empirical data. To do this, we assume that the other on-chip networks, such as logic, power and memory, have similar scaling characteristics as H-trees. This and the other assumptions given below are almost certainly overly simplistic, but they serve as a starting point. Even this preliminary theory reveals interesting features of power scaling in microprocessors.

The equation for active power is $P = aCV^2 f$ (Mezhiba & Friedman 2002), where P is active power, C is capacitance, a is an activity factor that represents the average percentage of C that is exercised, V is voltage and f is frequency. We use equation (3.3) to transform the power equation into a form that relates power to available data on die area, number of transistors, process size, voltage and clock frequency.

Capacitance is a complicated variable, but it can be estimated as $C = QA_{\mathrm{net}}$ (Ho 2003) where wire complexities (such as the m-factor, wire pitch and wire aspect ratio) are subsumed in Q. As an initial estimate, we assume that Q and a are constant factors, so that we can approximate $P \propto A_{\mathrm{net}} V^2 f$. Because this equation measures the active power of all wires on the chip, the formula for A_{net} must account for the footprint of all wire networks. By replacing the terminal units of the clock tree (N_{c}) with the terminal units of all wires (the number of transistors, N_{tr}), we assume that other on-chip networks scale similarly to the clock tree.

With these assumptions, we can estimate P using equation (3.3),

$$P \propto \lambda V^2 N_{\mathrm{tr}} A_{\mathrm{chip}}^{1/2} f. \tag{3.4}$$

3.2.3. Observations. Figure 4 shows a strong correlation between the prediction (equation (3.4)) and observed power across the 523 microprocessors. The r^2 is 0.79. A scaling exponent equal to 1 would indicate a close correspondence between theory and data. However, the scaling estimates are significantly lower than 1 (the OLS estimate is 0.47 and RMA is 0.53), indicating systematic deviations between the theoretical predictions and the data.

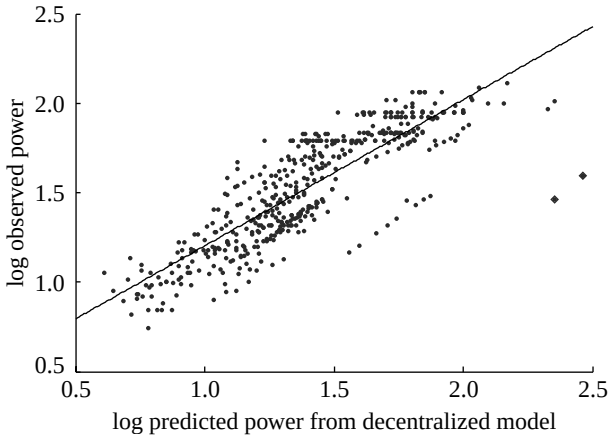

Figure 5. Power measured on microprocessors versus power predicted by equation (3.5) ($P \propto \lambda V^2 N_{\mathrm{tr}}^{1/2} A_{\mathrm{chip}}^{1/2} f$). The scaling exponent is close to 1 (OLS exponent is 0.82, the RMA exponent is 0.95, 95% CI is 0.76–0.99). Excluding the outliers in the lower right, the scaling exponent is indistinguishable from 1.

Thus, equation (3.4) does not provide an adequate explanation for power scaling. However, a regression analysis showed that active power is strongly correlated individually with each of λ, N_{tr} and A_{chip} across a variety of microprocessors (data not shown). The strong correlation suggests that these variables are relevant to predicting power, but the form of the equation is not correct. As a first step towards improving the theory, we performed a multiple regression to understand interactions among the variables. Notably, A_{chip} was an insignificant term in the multiple regression when interaction terms were included.

3.2.4. An alternative scaling prediction for power. Finally, we test whether an alternative equation for A_{net} gives predictions for power that are more consistent with the data. In the electronic supplementary material, we derive a prediction for the length of the H-tree from the central clock to any leaf of the network, and A_{net} is derived from that length. However, some networks may not be centralized like the H-tree. For example, in the logic network, wires are not routed from each transistor to a centralized control point. Instead, many wires connect transistors or cells that are located nearby. Intelligent place and route algorithms are used to minimize wire length by placing connected cells in close proximity.

Alternatively if we assume that the length of each wire scales as the distance between nearest components, then we can estimate the length of each wire from the density of components, $l \propto \rho_{\mathrm{c}}^{-1/2}$. Then we estimate $A_{\mathrm{net}} \propto Nl\lambda$ and given $P \propto A_{\mathrm{net}} V^2 f$ predict

$$P \propto \lambda V^2 N_{\mathrm{tr}}^{1/2} A_{\mathrm{chip}}^{1/2} f. \tag{3.5}$$

The only difference between equations (3.4) and (3.5) is that the exponent on N_{tr} changes from 1 to 1/2.

Observations. Figure 5 shows a close correspondence between the data and the alternative prediction (equation (3.5)). The slope is only slightly less than 1, (OLS estimated slope is 0.82, RMA is 0.95 with 95% CI between 0.76 and 0.99, r^2 is 0.75). The figure excludes the seven Intel chips from the 1970s and 1980s because they appear to be systematically different. Including them, the 95% CI is 0.74–0.93. The extreme outliers in the lower right corner are Intel Xeon Cascade chips that have unusually large L2 caches. Because our prediction depends on area and the L2 caches are less active than other regions of the chip, these chips dissipate less power than we predict. If those points are excluded from the regression, the r^2 increases and the scaling exponent increases so that it is indistinguishable from 1. We conclude that this decentralized model is more consistent with observed power than the centralized model.

3.3. Scaling in other information networks

MST suggests that delivering resources from a central source to all parts of an organism is a dominant design constraint and that efficient solutions to this problem have evolved via natural selection. We used a similar approach to characterize constraints in the design of H-trees and found that the footprint of the H-tree scales consistently with MST. However, power requirements appear to scale differently, consistent with a decentralized scaling model. The framework can be extended to other information networks.

Table 2 highlights seven network properties that have analogous function in biological networks (illustrated by the circulatory system) and three information networks: the brain, microprocessors and the Internet. In each system there is a network that delivers matter, energy or information to components distributed in physical space. All of these systems are designed to maximize some aspect of performance under time or power constraints. The optimizing process in biology is natural selection; in human built systems, it is engineering principles and economics, and in the case of the Internet, self-organization may also play a role. Our results suggest that engineered and natural systems have developed networks to achieve similar design goals under similar constraints.

We expect all of these systems to be subject to the fundamental constraint in equation 1.3 so that the size of the network increases faster than the number of components. Thus we expect $(D+1)/D$ scaling of network size with system size; when network size is constrained to be linear with system size, we also expect $1/D$ scaling of density and $D/(D+1)$ scaling of performance.

The mapping between physical network links and the material that flows through them is straightforward. All of these networks are built from components that obey the laws of physics, such as conservation of current or fluid flow through the network. These constraints can be accommodated in different ways, as illustrated by differences between the H-tree and the circulatory network. Other network properties, such as decentralization and connectivity patterns, may further alter the scaling constraints.

Additionally, each system has an interface between the network and the components it services. This last mile determines, for example, how oxygen travels from a capillary to a mitochondrion or how a clock signal moves from a clock buffer to a flip flop or latch in the

Table 2. Mapping between biological and information networks.

	circulatory system	brain	chips	Internet
components	capillaries/ mitochondria	neurons	transistors	hosts
links	blood vessels	dendrites, axons	wires	fibre
what flows	blood (oxygen, glucose)	action potentials	charge carriers (electrons)	packets (information)
the last mile	capillary to mitochondria	synapse	isochronic region (for clock trees)	local area network
network properties	hierarchical, fractal branching, space-filling, three dimensions	neural connectivity, space-filling, three dimensions	hierarchical, space-filling, 2.5 dimensions	hierarchical, space-filling, two dimensions
what is optimized	maximize metabolic rate, minimize transport costs and times	maximize information processing, minimize metabolic demand	maximize processing minimize power	maximize throughput minimize latency
fundamental principles	conservation of energy and matter	conservation of energy and matter	Kirchhoff's laws (conservation of current and voltage)	telegrapher's equation
driving process	natural selection	natural selection	engineering economics	engineering economics self-organization

isochronic region serviced by that clock buffer. MST does not yet address the role of the last mile, but it plays an important role in all of these systems and is an area where MST might be fruitfully extended.

Figure 6 shows scaling relationships for three systems from table 2 (scaling relationships for chips are shown in figures 2–5). As a preliminary test of the MST-inspired hypothesis, we can compare empirical data with the $D/(D+1)$ predictions of MST. The data illustrate different ways to incorporate network scaling into the system design. Three-dimensional organisms are subject to constraint D ($\mathrm{vol}_{\mathrm{net}} \propto \mathrm{vol}_{\mathrm{org}}$), so that the scaling of metabolism and the products of metabolism (such as biomass production) follow $B \propto M^{3/4}$. In three-dimensional brains, the data show $\mathrm{vol}_{\mathrm{net}} \propto N_{\mathrm{c}}^{4/3}$ (following equation 1.3) where $\mathrm{vol}_{\mathrm{net}}$ is approximated by the volume of white matter (axons), and N_{c} is represented by the volume of grey matter (neuron cell bodies). The Internet covers the two-dimensional surface of the Earth and shows 2/3 power scaling of bandwidth, suggesting that the rate of communication per host slows by $1/(D+1)$ scaling as the number of hosts increases. In this case, the density of hosts is not reduced, but the proportion of processing that uses the network is reduced. Figure 3 shows that the area of the network scales faster than the area of the chip, necessitating additional metal layers. These are very different networks, but they face a common constraint: resources are distributed through a network that scales superlinearly with the number of components to which it delivers energy or information.

4. DISCUSSION

We extended the MST framework to characterize the scaling behaviour of microprocessors, particularly focusing on the clock trees. The two most important extensions were relaxing the assumption of invariant terminal units (allowing lengths of terminal units to vary with their density) and accounting for decentralized networks. The H-tree is an example of an information network that is geometrically and functionally similar to vascular networks. It is a centralized, hierarchical distribution network designed to optimize power and performance within a constrained physical space. In the H-tree, as in vascular networks, the size of the network grows faster than the number of components it connects.

While a common scaling framework describes H-trees and cardiovascular systems, there are important differences in how networks scale in the two systems. The widths of terminal wires in H-trees are determined by process size (λ) and lengths of terminal wires change with their density (and therefore with the area of isochronic regions and frequency of the clock), but they are hypothesized to be approximately invariant in organisms. When terminal wires are allowed to vary, the predicted clock tree footprint (A_{net}) changes from equation (3.2) to (3.3). However, these equations do not account for an important innovation in clock trees, which allows wire width and power to be reduced, violating the assumption of width-preserving branching. This innovation, the use of *repeaters* to amplify the clock signal, produces more efficient clock tree designs. Accounting for repeaters in the model is left for future work.

The systems also accommodate super-linear network scaling differently. Organisms decrease component density in order to keep network volume linear with organism volume, while integrated circuits keep component density constant but use the third dimension to hold excess wire area. Our predictions for H-tree scaling are consistent with both the simulations and empirical measurements of metal layers (figures 2 and 3).

We hypothesized that the excess wire associated with larger chips would result in a predictable increase of active power with A_{chip}. Although there is a positive correlation between our predictions and the observations, there are systematic deviations, and A_{chip} is

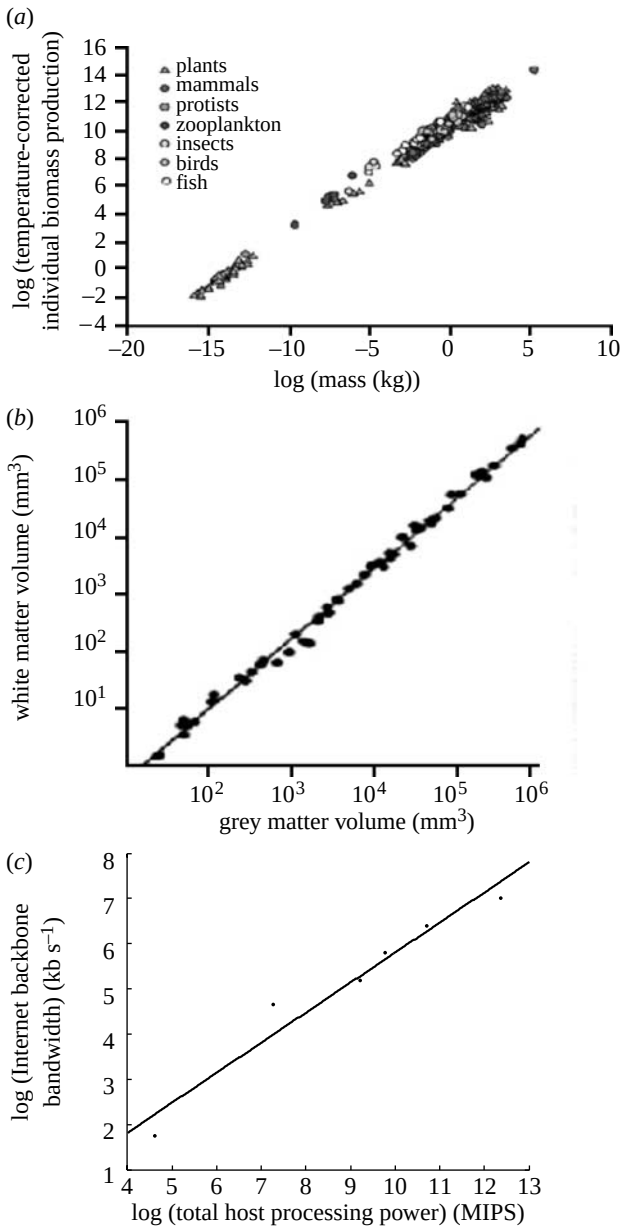

Figure 6. Scaling relationships in biological and computational systems. The slope of the fitted regression line on log-transformed axes gives the scaling exponent, b, for (*a*); the rate of biomass production due to growth and reproduction as a function of body mass in different groups of organisms (observed $b = 0.76$, predicted $b = 3/4$, figure from Ernest *et al.* 2003); (*b*) volume of grey matter (neurons) as a function of white matter (axons) in mammalian brains (observed $b = 1.23$, predicted $b = 4/3$, figure from Harrison *et al.* 2002); (*c*) Internet backbone bandwidth as a function of the processing power of the Internet hosts (observed $b = 0.66$, predicted $b = 2/3$, data from Moore (2003), Internet Systems Consortium (2008, http://www.isc.org/index.pl?/ops/ds/host-count-history.php) and Zakon (2006)).

not an important determinant of power consumption when other variables are taken into account. The deviations may occur because we subsumed much of the complexity of power analysis into scaling constants. Alternatively, the assumption that other on-chip networks such as logic, power and memory have scaling properties similar to those of H-trees may be incorrect. These networks have different functions and perhaps topologies that scale differently. Finally, we note that area and density are each related, but in different ways, to aspects of the fabrication process that are not incorporated directly into the model. Area directly affects yield, which is the expected fraction of fabricated chips that are operational. Hence area is dependent upon material and fabrication quality. Density is influenced by λ and the number of available metal layers associated with the fabrication process.

Given the strong correlation between power and the variables in equation (3.4), we tested an alternative model of network scaling. We found that a model that assumed a completely decentralized network in which each component was connected only to its nearest neighbour was a better predictor of power than the centralized model. In this alternative model, the footprint of the network depends on the density of components rather than the size of the system. Thus, active power scales as though the relevant networks (perhaps primarily the logic network) are decentralized, but the metal layers scale like the network footprint (perhaps dominated by global networks such as the H-tree) scales consistently with MST predictions for centralized networks. The relationships between power, wire scaling and component density are of practical significance because all three properties are key to chip performance (Mezhiba & Friedman 2002). More research is needed to understand these scaling relationships.

Equations (3.1)–(3.5) are first steps towards a scaling theory for information networks. Our goal is to develop a predictive theory for microprocessors that is derived from first principles. Such a theory would be an important contribution for several reasons. First, it would provide engineers a quantitative metric for their designs in terms of a theoretical ideal. This would inform decisions about when to continue optimizing a current design strategy versus looking for radical innovations. Because the predictions of the MST-inspired scaling theory arise from structural rather than behavioural properties, this approach could potentially provide a new tool for predicting performance. Currently, new designs are evaluated using expensive simulations. Because the simulations are so time-consuming, it is not feasible to test multitudes of designs in detail, nor is it feasible to obtain precise results. A theory that can predict hard-to-measure properties, such as power, from easy-to-measure properties, such as process width and transistor count, would be a welcome contribution, even if the predictions are initially somewhat crude. However, it is important to respect the limits of precision that are inherent in an order-of-magnitude scaling theory. The current formulation is not detailed enough to predict the exact power requirements of any specific design.

The scaling approach could be useful for characterizing other on-chip networks. For example, H-trees are used in dynamic random access memory (Jouppi 2006), asynchronous chips (Takamura & Fukasaku 1997) and field programmable gate arrays (Zhu & Wong 1997). Hierarchical designs that attempt to mitigate geometric scaling are evident in cache hierarchy and power networks. Multi-core chips address the clock tree scaling problem by reducing

the distance that any signal has to travel to within a given core. However, multi-core architectures introduce new networking problems, arising from the need to coordinate activities among the different cores. 'Networks on chip' have been proposed for distributing control and data messages among cores and other components. Further extensions of MST could help predict the limits of this design strategy.

Extending scaling analysis to other networks reveals additional commonalities among disparate systems. For example, brains and microprocessors allow wires to scale in a different dimension than the components to accommodate super-linear wire scaling. In brains, neuron cell bodies (grey matter) cover the two-dimensional surface of the brain, while axons and dendrites (white matter) fill the interior three-dimensional volume, similar to the engineering solution of using a two-dimensional surface to hold transistors and metal layers that extend into the third dimension to hold wires. Another common feature across these networks is the interface between the network and the components (the last mile). In many cases, resource flow through the last mile is considerably slower than flow through the network. For example, blood is pumped quickly through large arteries, but oxygen slows to diffusive speeds to exit capillaries and enter cells; wire speeds on the silicon layer are slower than that on metal layers; wireless connections to the Internet are slower than backbone connections. The difference between speed through the network and that through the last mile probably has significant impact on system performance and could perhaps be described by extending MST.

An important issue for further investigation is how the degree of centralization affects network scaling. Some networks are highly centralized (e.g. the cardiovascular system and the H-tree), while others (e.g. brains and the Internet) have no central controller or repository of information. Multi-core architectures are now the focus of the microprocessor industry; in these devices there is centralization within each core but each core functions autonomously. These decentralized systems may exhibit different scaling exponents. Even in the case of H-trees, there has been a trend towards decentralization by inserting repeaters at branch points. Repeaters allow a signal to be sent from the central clock and amplified as needed to reach components. This innovation is primarily motivated by signal integrity issues that become more problematic at smaller process geometries and may reduce power requirements and alter the prediction shown in equation (3.3). Our initial predictions for power scaling on microprocessors assumed that the wire networks that consume most of the power are centralized like the H-tree; that assumption is untested and may explain why equation (3.4) does not match observations. We tested an alternative scaling model in which wires connect only local components, and that model was more consistent with empirical data. Thus, it appears that network scaling on microprocessors depends on both the density of components and the area of the chip. This is similar to observations in urban road networks: the area or 'lane-miles' occupied by these road networks depends on both the density of the population and the physical area of the city (Samaniego & Moses 2008). Accounting for partial decentralization in networks is a fruitful area for future research.

Natural selection and engineering have independently converged on similar network designs. Even though these processes are distinct, they rely on competition and optimization within some constraint space. We have shown one case of how evolution and engineering have discovered similar network design principles; we think there are others. In comparing organisms with engineered systems, it is important to note that organisms are the result of billions of years of evolution, while chip designs have been optimized by engineers for only a few decades. In these engineered systems, some aspects are locked in to ensure backward compatibility (e.g. the application programmer interface), but other components are continually modified (e.g. materials and process technology). This contrasts with biological systems in which evolution locked into certain design elements long ago. Eventually, as component features such as process size reach their physical limits, and as the limits of this architectural regime are approached, we expect either radical innovations or engineering solutions that look more biological.

The authors thank anonymous reviewers for constructive comments on earlier versions of this manuscript and George Bezerra who contributed valuable insights and made suggestions. We also gratefully acknowledge the support of NSF grant CCF 0621900 and the Santa Fe Institute. M.E.M. and J.H.B. acknowledge partial support from NSF DEB-0083422 and Los Alamos grant W-7405-EN6-36; M.E.M. acknowledges partial support from NIH P20 RR-018754; S.F. acknowledges partial support from NSF CCR-0331580 and CCR-0311686 and AFOSR FA9550-07-1-0532; and A.L.D. and M.A.L. acknowledge funding from NSF 162 941 and NSF 0430063 and thank HP Laboratories for its support.

REFERENCES

Banavar, J. R., Maritan, A. & Rinaldo, A. 1999 Size and form in efficient transportation networks. *Nature* **399**, 130–132. (doi:10.1038/20144)

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. & West, G. B. 2004 Toward a metabolic theory of ecology. *Ecology* **85**, 1771–1789. (doi:10.1890/03-9000)

Brown, J. H., West, G. B. & Enquist, B. J. 2005 Yes West, Brown and Enquist's model of allometric scaling is both mathematically correct and biologically relevant. *Funct. Ecol.* **19**, 735–738. (doi:10.1111/j.1365-2435.2005.01022.x)

Dawson, T. 2003 Scaling laws for capillary vessels of mammals at rest and in exercise. *Proc. R. Soc. B* **270**, 755–763. (doi:10.1098/rspb.2002.2304)

Deng, Y. & Maly, W. 2004 2.5D System integration: a design driven system implementation schema. In *Proc. Asia South Pacific Design Automation Conf.*, pp. 450–455. See http://doi.ieeecomputersociety.org/10.1109/ASPDAC.2004.2

Ernest, S. K. M. *et al.* 2003 Thermodynamic and metabolic effects on the scaling of production and population energy use. *Ecol. Lett.* **6**, 990. (doi:10.1046/j.1461-0248.2003.00526.x)

Etienne, R. S., Apol, M. E. F. & Olff, H. 2006 Demystifying the West, Brown & Enquist model of the allometry of metabolism. *Funct. Ecol.* **20**, 394–399. (doi:10.1111/j.1365-2435.2006.01136.x)

Friedman, E. G. 2001 Clock distribution networks in synchronous digital integrated circuits. *Proc. IEEE* **89**, 665–692. (doi:10.1109/5.929649)

Glazier, D. S. 2005 Beyond the '3/4-power law': variation in the intra- and interspecific scaling of metabolic rate in animals. *Biol. Rev.* **80**, 611. (doi:10.1017/S1464793105006834)

Harrison, K., Hof, P. & Wang, S. 2002 Scaling laws in the mammalian neocortex: does form provide clues to function? *J. Neurocytol.* **31**, 289–298. (doi:10.1023/A:1024178127195)

Ho, R. 2003 On-chip wires: scaling and efficiency. PhD thesis, Stanford University.

Jouppi, N. 2006 Cacti v. 4.0, Hewlett-Packard Laboratories technical report, HPL-2006-86.

Kleiber, M. 1932 Body size and metabolism. *Hilgardia* **6**, 315–353.

Kozlowski, J. & Konarzewski, M. 2004 Is West, Brown and Enquist's model of allometric scaling mathematically correct and biologically relevant? *Ecology* **18**, 283–289. (doi:10.1111/j.0269-8463.2004.00830.x)

Makarieva, A. M., Gorshkov, V. G. & Li, B.-L. 2005*a* Revising the distributive networks models of West, Brown and Enquist (1997) and Banavar, Maritan and Rinaldo (1999): metabolic inequity of living tissues provides clues for the observed allometric scaling rules. *J. Theor. Biol.* **237**, 291–301. (doi:10.1016/j.jtbi.2005.04.016)

Makarieva, A. M., Gorshkov, V. G. & Li, B.-L. 2005*b* Biochemical universality of living matter and its metabolic implications. *Funct. Ecol.* **19**, 547–557. (doi:10.1111/j.1365-2435.2005.01005.x)

Mead, C. & Conway, L. 1979 *Introduction to VLSI systems.* Boston, MA: Addison-Wesley.

Mezhiba, A. & Friedman, E. 2002 Trade-offs in CMOS VLSI circuits. In *Trade-offs in analog circuit design: the designer's companion* (eds C. Toumazou, G. S. Moschytz & B. Gilbert), pp. 75–114. Boston, MA: Kluwer Academic Publishers.

Moore, G. 2003 *IEEE Int. Solid-State Circuits Conference, Key Note Address.*

Moses, M. E., Hou, C., Woodruff, W. H., West, G. B., Nekola, J. C., Zuo, W. & Brown, J. H. 2008 Revisiting a model of ontogenetic growth: estimating model parameters from theory and data. *Am. Nat.* **171**, 632–645. (doi:10.1086/587073)

O'Connor, M. P., Agosta, S. J., Hansen, F., Kemp, S. J., Sieg, A. E., McNair, J. N. & Dunham, A. E. 2007 Phylogeny, regression, and the allometry of physiological traits. *Am. Nat.* **170**, 431–442. (doi:10.1086/519459)

Peters, R. 1983 *The ecological implications of body size.* Cambridge, UK: Cambridge University Press.

Samaniego, H. & Moses, M. E. 2008 Cities as organisms: allometric scaling of urban road networks. *J. Trans. Land Use* **1**, 21–39.

Savage, V. M., Gillooly, J. F., Woodruff, W. H., West, G. B., Allen, A. P., Enquist, B. J. & Brown, J. H. 2004 The predominance of quarter-power scaling in biology. *Funct. Ecol.* **18**, 257–282. (doi:10.1111/j.0269-8463.2004.00856.x)

Schmidt-Nielsen, K. 1984 *Scaling: why is animal size so important?* New York, NY: Cambridge University Press.

Stroobandt, D. 2001 *A priori wire length estimates for digital design.* Boston, MA: Kluwer Academic.

Takamura, A. & Fukasaku, I. 1997 TITAC-2: an asynchronous 32-bit microprocessor based on scalable-delay-insensitive model. In *Proc. Int. Conf. Computer Design (ICCD '97), 12–15 October 1997*, p.288. See http://doi.ieeecomputersociety.org/10.1109/ICCD.1997.628881

Warton, D. I., Wright, I. J., Falster, D. S. & Westoby, M. 2006 Bivariate line-fitting methods for allometry. *Biol. Rev.* **81**, 259–291. (doi:10.1017/S1464793106007007)

West, G. B., Brown, J. H. & Enquist, B. J. 1997 A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126. (doi:10.1126/science.276.5309.122)

White, C. R. & Seymour, R. S. 2003 Mammalian basal metabolic rate is proportional to body mass $^{2/3}$. *Proc. Natl Acad. Sci. USA* **100**, 4046–4049. (doi:10.1073/pnas.0436428100)

Zakon, R. H. 2006 Hobes internet timeline, v. 8.2. See http://www.zakon.org/robert/internet/timeline/(2006).

Zhu, K. & Wong, D. 1997 Clock skew minimization during FPGA placement. *IEEE Trans. Comput. Aided Design Integr. Circuits Syst.* **16**, 376–385. (doi:10.1109/43.602474)