IEEE Access
practical innovations : open solutions

Received June 5, 2013, accepted June 24, 2013, date of publication July 3, 2013, date of current version July 16, 2013.

Digital Object Identifier 10.1109/ACCESS.2013.2272175

A Card Stack Model to Elucidate Key Challenges in the Development of Future Generation Supercomputers

WATARU NAKAYAMA (Life Fellow, IEEE)
ThermTech International, Kanagawa 255-0004, Japan (watnakayama@aol.com)

This work was supported by the Industry/Academia Project "Advancement of Reliability Science and Engineering for Electronic Equipment" organized under the auspices of Japan Society of Mechanical Engineers.

ABSTRACT This paper intends to elucidate challenges in some aspects of the hardware design of future generation computers. We use a system model, a stack of integrated circuit cards cooled by a dielectric coolant (FC77). A set of equations is developed to describe the relationships between the system throughput, the volume, the power consumption, and those concerning the details of internal organization such as signal and power line dimensions and coolant path width. The calculated values of throughput, volume, and power are projected on a state point in a graph of the figures-of-merit pair, the computational density, and the computational efficiency. By manipulating the empirical parameters imbedded in the model, the state point is steered to follow the evolutionary line that runs through the points corresponding to the existing supercomputers of several generations. Then, calculation is extended on state points for future prospective computers with target system throughputs. The results point to the needs for research and development effort on thermal management and materials development. As for thermal management of exa- and zeta-scale computers, we need to refocus heat transfer research. Coolant channels will have very large length-to-width ratios (several thousand), while the heat flux on the channel surface is quite low. Micro-fluidics to guarantee stable coolant flow in such long micro-channels will be of primary importance in place of the means to deal with high heat flux. We also need to develop novel materials for signal transmission lines and cooling, particularly in the development of zeta-scale computers.

INDEX TERMS Computational efficiency, computational density, dielectric coolant, immersion cooling, hardware, supercomputer, system-level modeling.

I. INTRODUCTION

One of the major technological challenges facing the science and engineering of the 21st century is the development of supercomputers that deliver exa-scale throughput and beyond. As of 2012, the throughput of supercomputers has exceeded 10 peta (10×10^{15})–FLOPS [1]. To advance the computational capacity further to exa (10^{18}) -FLOPS we must curb explosive growth of system volume and power consumption. The U.S. Defense Advanced Research Projects Agency (DARPA) sets the target figures for an exa-scale computer such that the system is housed in 500 ordinary-sized racks and the power requirement at 20 MW [2]. A study based on the projection of current technology trends forecasts that an exa-scale system could be contained in 583 racks, but the power would be more like 500 MW [2]. For zeta (10^{21})-scale computing the image of hardware of reasonable physical volume and power consumption is hard to envision.

In our attempt to develop the hardware image of a future prospective supercomputer, we need to work on the basis of the following premise. For the processing performance of a computer the flight time of signals between circuit elements is the controlling factor. Time of flight is proportional to the length of a communication line linking the circuit elements. In a computer involving a huge number of circuit elements, routing of communication lines poses a formidable challenge. The lengths of all communication lines must be contained in a certain acceptable range. The requirement for reduced routing lengths necessarily leads to packing circuits in a three

dimensional space. Today, three-dimensional packaging of electronic circuits is already in progress in small systems. Stacking memory chips saves the area on the printed circuit board, and reduces the communication distance between the logic chip and memory cells. The next step is stacking memory chips on top of the logic chip, which further cuts the distance between them. Stacking logic chips seems the logical extension of this trend; however, there are some issues we need to address in the work on this scheme.

The first issue of concern for stacked chips is thermal management. Heat removal from stacked-chip systems has been a topic of active research in recent years. Common to almost all of the existing studies is the assumption of high heat flux on the chip's surface; for example, 135 W/cm^2 [3] and 390 W/cm^2 [4]. To deal with high heat flux water is chosen as a coolant [3]–[5]. Such premise of the study, however, needs cautionary examination. We note that the direct water-cooling has not been materialized in actual machines in the past. Indirect water cooling has been used in large-scale computers, but it requires a space for conduction devices which is hardly affordable in a chip stack. The second but equally disconcerting issue is the uncertainty in regard to where such high-power chip stacks will find applications [6]. Construction of a large system by assembling a number of high-powered chip stacks seems a remote possibility.

An alternative to water cooling is immersion cooling by dielectric coolant; however, the dielectric coolant has low thermal conductivity. We will not be allowed to increase the heat dissipation rate on the chip's surface at the rate we have experienced to this day. We may be driven back to the design concept of CRAY supercomputers of the 1970s and 80s, where the computing performance is dictated by the routing lengths of wires for signal transmission within the system. Trimming of wires to minimal lengths necessarily leads to dense packing of circuit elements in a three-dimensional space.

In our attempt to envision challenges in the construction of future supercomputers we assume three-dimensional systems cooled by dielectric coolant. We need a model which embodies three-dimensionality of the system, yet is simple enough to allow us to develop a system of equations that describe the relationships between various parameters. The parameters of our interest are the system's processing throughput, the volume, the power consumption, and those parameters concerning the details of internal organization such as the spaces occupied by signal and power lines and coolant flow. The model of our choice is composed of integrated circuit cards placed in stack; it has communication lines on the sides of the stack, and coolant paths between the cards. Those components in actual computers, such as chips, packages, and wiring boards, are not explicitly represented in the model. The reduced architecture model is necessary to maintain the complexity of analysis at a reasonable level.

We also need a guide map in which we trace the evolution of supercomputers of the past and the present generations, then, locate the points of exa-scale and zeta-scale computers on the extension of the evolution curve. We find the framework to create an evolution map in the paper by Ruch et al. [7]. The graph in Ruch et al. [7] has a pair of the figures of merit on its coordinates, the computational density and the computational efficiency. They plotted the data from a wide range of computers in the graph, and showed that the data fall close to the diagonal line, and the technology evolution is represented by the shift of data points towards the upper right corner of the graph, indicating the achievement of higher densities and efficiencies. In the present study the data from several examples of the past and existing supercomputers are used to create foothold points in the graph, and from there the projection of requirements for future technologies is attempted.

As already mentioned above, the model used in the present study is stripped of some structural details of actual computers. Also left outside the scope of the study are the architectural aspects of computer design. Dongarra [8] reviewed the development of high-performance computing technology up to the mid-2000s. Coteus et al. [9] discussed the importance of memory management technology for realization of exa-scale systems. In a more detailed version of the model we may consider job streams in the system's physical space, particularly, those involving processing and memory circuits. In the present model those architectural details are 'homogenized.'

The paper is organized as follows. Section II describes the physical model, and explains the ingredients of mathematical modeling. Section III describes the calculation steps, and specifies the values of the relevant parameters used in the case studies. Section IV reports the results of the case studies and discussions. Section V concludes the paper.

II. MODEL DESCRIPTION

The system model is a stack of cards shown in Fig. 1. The card carries circuit elements that perform elementary logic calculations. Fig. 1 illustrates how processing element A is connected to element B on a different card; signal from A is led through a metal line (xy-line) to the card edge, then, sent to the edge of the card carrying B through an optical line (z-line), and to B on a metal line on the card. The side tiles on the two sides of the stack accommodate z-lines. The cards are arranged with the spacing for coolant flow (d_f); the sum of d_f and the card thickness (d_s) is the placement pitch of cards (d_p). We suppose square cards having a side length L, and the stack extends to a length L_z. The system has N_S elements partitioned to M cards, so that each card accommodates $N_C = N_S/M$ elements. In pursuit of ever larger scale computing we increase N_S. The number of cards, M, hence, the stack length L_z, depends on the partition policy; that is, in what proportion we divide N_S elements to M cards. In general, there are constraints on the implementation of circuit elements on a single card. One of the constraints is the economy of manufacturing a large area card; in actual terms this means the constraint on the size of a chip, a wafer, and a PCB. The other constraint is on the density of circuit elements on a card; the looming physical limit to the extension of Moore's law manifests this

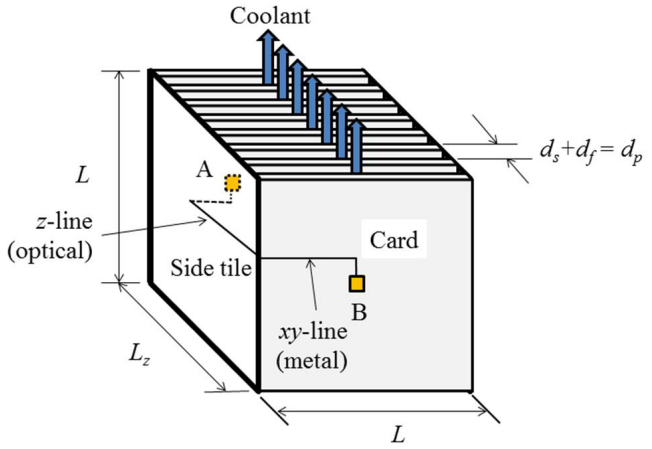

FIGURE 1. System model: Cards carrying circuit elements are placed in stack with spatial allowance for coolant flow. An illustration includes communication lines between circuit elements A and B on different cards. In the model, actual hardware components such as chips, packages, and printed circuit boards are 'homogenized' in the card.

constraint. Under these constraints, M, hence, L_z, increases with increasing N_S. With elongation of the stack the signal transmission time between the cards increases, making negative impact on the system performance. Furthermore, the space for coolant flow adds an overhead on the stack length; more overhead results as we accommodate more cards in the stack. The system's physical construction, processing performance, and power consumption are coupled through several mechanisms. The present model has these mechanisms as its key parts, which we formulate in the following sub-sections.

A. AVERAGE LINE LENGTH

The average line length serves as a measure for signal transmission delay within the system. It also serves as a measure for power consumption by the circuit element, as the capacitance to electrically charge a line is a major source of power consumption by an element in sending signal to other elements. We use the Donath formula for the average line length on the card [10]. In an abbreviated form for $N_C >> 1$ and $1/2 < p < 1$, it is

$$\bar{R}_C = \frac{14}{9} \cdot \frac{1-4^{p-1}}{4^{p-1/2}-1} N_C^{(2p-1)/2} \tag{1}$$

where p is the Rent exponent. $\bar{R}_C$ is non-dimensional, measured in unit of $L/\sqrt{N_C} \equiv d_e$, so that the dimensional average line length is $\bar{l}_C = \bar{R}_C d_e$.

The average length of lines connecting elements on different cards is derived in a similar way that produced (1). The author's earlier paper [11] reports the derivation, but some corrections are introduced in the formula shown below. We assume that elements A and B in Fig. 1 are both in the left or right half of the card. Where this is not the case, we suppose that A, for example in the left half, communicates the image of B in the right half, and signal travels from the image of B to the actual B through a z-line. Communication between A in the left half and the image of B in the right half on the same card is counted as on-card communication, so that the line length between them is accounted for by (1). Hence, as far as longitudinal communications are concerned, we have symmetry of line routing with respect to a mid-plane cutting through the stack, and consider a stack of half width ($L/2$). We apply hierarchical partition to this half-wide stack. The stack is divided into subsets; at level-k partition each subset has 2^k cards. Since each card has N_C /2 elements on its half area, a level-k subset has $2^{k-1}N_C$ elements. According to the Rent rule the number of z-lines emerging from this subset of cards is

$$T_k = A\left(2^{k-1}N_C\right)^p \tag{2}$$

where A is the Rent constant. There are $N_S/2^k N_C$ subsets, and we write the number of z-lines interconnecting subsets of $2^{k-1}N_C$ as $\alpha T_k\left(N_S/2^k N_C\right)$, where α is a factor. This includes the number of interconnections for subsets finer than level k. To calculate the number of interconnects belonging to level k, we subtract components belonging to levels $(k+1)$ and finer.

$$\begin{aligned} m_k &= \alpha T_k\left(N_S/2^k N_C\right) - \alpha T_{k+1}\left(N_S/2^{k+1}N_C\right) \\ &= \alpha A N_S N_C^{p-1} 2^{k(p-1)-p}\left(1-2^{p-1}\right). \end{aligned} \tag{3}$$

The average length of z-lines is

$$\bar{R}_z = \sum_{k=0}^{K-1} m_k r_{zk} \Big/ \sum_{k=0}^{K-1} m_k \tag{4}$$

where $K = \log_2\left(N_S/N_C\right) = \log_2 M$, and r_{zk} is the non-dimensional distance between neighboring subsets of level-k, which is 2^k. The reference length for r_{zk}, hence, for $\bar{R}_z$, is the placement pitch of cards d_p. After some manipulation we obtain, for $M >>1$ and $p <1$,

$$\bar{R}_z = \frac{1-2^{p-1}}{2^p-1} \cdot M^p. \tag{5}$$

The dimensional average length of z-lines is $\bar{l}_z = \bar{R}_z d_p$.

In addition to the average z-line length we need to take into account the line lengths from elements to the card edges and the relative displacements between the projected images of elements on the card. Following the procedure used in deriving (1) we obtain the contribution of these on-card lines as, for $N_C >>1$

$$\Delta\bar{R}_z = \frac{5}{6} N_C^{1/2}. \tag{6}$$

In the dimensional form this overhead is $\Delta\bar{l}_C = \Delta\bar{R}_z d_e = \left(5/6\right) L$.

The system-level average line length is the weighted average of the above derived components. For simplicity we write the system line length setting $d_e = d_p$ as

$$\bar{R}_{sys} = \frac{M \cdot \Sigma_n \cdot \bar{R}_C + \Sigma_m \cdot \left(\bar{R}_z + \Delta\bar{R}_z\right)}{M \cdot \Sigma_n + \Sigma_m}. \tag{7}$$

The weighting factor for the z-line routing is $\Sigma_m = \sum_{k=0}^{K-1} m_k$, where m_k is substituted from (3). The weighting

factor for the on-card routing, Σ_n, is a similar summation of the interconnection pairs on the card. We do not elaborate a more detailed form of (7), because we will use it only as a background material in the next section. Also to be noted is that the assumption of $d_e = d_p$ is used only in calculation of $\bar{R}_{sys}$; hence, the main body of analysis does not employ this assumption.

The routing of transmission lines in a three-dimensional computing system has been studied by many investigators. In most of the existing models z-lines are not constrained in the side tiles but given equal probabilities for routing as xy-lines, for example, [12]–[14]. Hence, the formulas of line length calculation reported in the literature are different from the one derived in the present study.

B. PARTITION POLICY

One may assume that, in optimum partition, N_S elements are partitioned into M cards so as to minimize $\bar{R}_{sys}$ of (7). We, however, do not base our partition policy on minimizing $\bar{R}_{sys}$ of (7) for two reasons. First, the curve of $\bar{R}_{sys}$ versus M is flat near the minimum point, M_{Rmin}. Second, the cost of assembling a stack has to be weighed against the cost of implementing circuit elements on a 2D plane. Denoting the cost of implementing a circuit element is y_e, and that of interconnecting the cards is y_c per card, we write the total cost of assembling the system as

$$Y = y_e N_C + y_c M = Y_0 / M + y_c M \tag{8}$$

where Y_0 is the cost of implementing all N_S elements on a single card.

Y of (8) becomes minimum at $M_{Y\min} = \sqrt{Y_0 / y_c}$. The cost ratio is difficult to estimate, particularly for future systems. For a rule-of-thumb estimation we use the data belonging to the current generation of computers. Assuming a system size $N_S = 10^8$, we refer to the ratio of the cost of a CPU board to that of connectors. Then, we substitute $Y_0/y_c \approx 2000$ and find $M_{Ymin} \approx 45$. Note that we use the cost factor only as an implicit reference in defining our partition policy. We choose a point near this value of M_{Ymin}, and make M grow with increasing N_S at the rate along the M_{Rmin} (true minimum point) -versus-$\bar{R}_{sys}$ relationship. Thus, we define the partition policy as

$$M = 0.04 N_S^{0.4}. \tag{9}$$

C. PROCESSING PERFORMANCE

The system's processing performance (throughput) is defined in terms of operations per second. For massively parallel processing the number of parallel communication lines and the flight time of signals are the determinants of the throughput. In the stacked system the line length and the signal flight time are defined separately for those circuits on the cards and the z-lines along the stack. We suppose that the primary constraint on the system performance is posed by signal flows along the stack. On the card edge there are N_{edge} ports for z-lines, and using Rent's rule we write

$$N_{edge} = \frac{1}{2} A \cdot N_C{}^p \tag{10}$$

where 2 in the denominator accounts for the symmetry of the circuit deployment on the cards.

There are in total M card edges; hence, $M \times N_{edge}$ of z-lines provides a capacity of parallel communications. We suppose that a fraction of them are active during time interval τ_z, and write the system throughput as

$$S_{TP} = \frac{A_C}{2\tau_z} M \cdot N_C{}^p \tag{11}$$

where A_C is the product of Rent's constant A and an activity factor. The time interval τ_z is the sum of the flight time of signal in an optical z-line and the time overhead on the optical modulator and receiver pair (τ_{zd}).

$$\tau_z = \frac{\bar{l}_z}{c_0} + \tau_{zd} \tag{12}$$

where $\bar{l}_z = \bar{R}_z d_p$ as defined earlier, and $c_0 = 3 \times 10^8$ m/s (speed of light).

D. POWER CONSUMPTION

A major part of power consumption in actual computer occurs on chips. On the chip, power is consumed by logic gating and signal transmission between the gates. With increasing circuit integration the driver circuits to charge the interconnect lines become dominant power consumer [22]. The chip also dissipates power to send signals to other chips through lines in the wiring substrate. In the present model, we capture chips and wiring substrates implicitly in a 'homogenized' complex of driver circuits and interconnect lines. Such modeling is employed in other system-level modeling as well (for example, [12]). We also take into account Joule heating of power/ground lines, the model of which is detailed in Appendix. The following derivation of equations needs to be followed with understanding of the above modeling concept.

The energy to charge a line of length l_e is written as

$$E_e = \frac{1}{2} \varepsilon \cdot l_e \cdot V_D^2 \tag{13}$$

where ε is the dielectric constant of the matrix in which the lines are embedded, V_D is the power line voltage, and l_e is the average length of active lines on the card calculated from

$$l_e = \left\{ 1 - (\gamma_A N_C)^{-(1-p)} \right\} \cdot \bar{l}_C + (\gamma_A N_C)^{-(1-p)} \cdot \Delta\bar{l}_C. \tag{14}$$

Equation (14) is derived on the following model. We suppose a fraction (γ_A) of N_C elements is active at any instant. Among them, those elements that participate in on-card processing are in the number $\gamma_A N_C - (\gamma_A N_C)^p$, and those sending signals to the card edge for longitudinal communications are $(\gamma_A N_C)^p$. These numbers are divided by $\gamma_A N_C$, and the ratios are used as the weighting factors for the average line lengths $\bar{l}_C$ (internal processing lines) and $\Delta\bar{l}_C$ (lines to the card edges).

The element electric power is

$$P_e = \frac{n_l}{\tau_C} E_e \tag{15}$$

where $n_l = 1/(1-p)$ is the number of lines attached to the element, and τ_C is the delay time on the on-card metal line. The delay time has two components; RC delay and transmission delay. For estimation of power consumption we use the formula of RC delay

$$\tau_C = \rho \cdot \varepsilon \cdot \frac{\bar{l}_C^2}{A_l} \tag{16}$$

where ρ is the signal line resistivity, and A_l is the cross sectional area of signal line. In deriving (16) we simplify the original equation for RC delay ([15], p. 206~207) by setting the product of the height of the metal line cross section and the underlying dielectric thickness equal to A_l. We further introduce the scaling factor, S_l, by which

$$A_l = S_l \cdot \bar{l}_C^2 \tag{17}$$

The element thermal power includes Joule heating of power and ground lines which is represented by a factor f_{JH}; hence,

$$P_{eH} = f_{JH} \cdot P_e \tag{18}$$

The derivation of f_{JH} is given in Appendix.

The system heat is the product of the element thermal power and the number of active elements; hence,

$$Q_{sys} = P_{eH} \cdot \gamma_A \cdot N_S \tag{19}$$

Heat per card is

$$Q_C = Q_{sys}/M. \tag{20}$$

E. SYSTEM VOLUME

The system volume is calculated from

$$V_{sys} = M \cdot L^2 \cdot d_p \tag{21}$$

The card pitch contains the following components: In derivation of the expression for the thickness components we assume that the power and signal lines are accommodated in a cross sectional area of the card the vertical height of which is fixed at L (card side length). The power lines supply electric power to the card, and the required thickness at the card edge is

$$d_{sp} = \frac{39 + 16\sqrt{6}}{30} \cdot \frac{\alpha_R \cdot \rho_p}{f_{JH} \cdot V_D^2} \cdot Q_C \tag{22}$$

where α_R is a factor concerning the voltage drop along the power-ground lines. The derivation of (22) is described in Appendix.

The signal transmission lines are embedded in dielectric matrix, and we assume that a line of cross sectional area A_l needs twice that area at the card edge to accommodate insulation by dielectric. The number of signal lines at the card edge is $A \cdot N_C^p$, where A is the Rent constant; hence, the thickness required to accommodate signal lines is written as

$$d_{ss} = 2\frac{A \cdot A_l}{L} N_C^p \tag{23}$$

In addition to d_{sp} and d_{ss} we introduce for the sake of generality an extraneous solid part of the card, d_{s0}; hence, we write

$$d_p = d_{sp} + d_{ss} + d_{s0} + d_f. \tag{24}$$

F. HEAT TRANSFER

We assume the flow of single-phase coolant in the channels between the cards. The coolant flow rate is constrained by either one of the following factors; the pressure drop between the inlet and the outlet (Δp), and the maximum allowable coolant velocity (V_{fmax}). Where the pressure drop constraint is effective, the coolant velocity (V_f) is calculated by

$$V_f = \Delta p \cdot \frac{d_f^2}{12\rho_f \nu_f L} \tag{25}$$

when the Reynolds number $Re \equiv V_f (2d_f)/\nu_f$ is below 2500 (laminar flow), or

$$V_f = \left(\frac{2^{9/4}}{0.316} \cdot \frac{1}{\rho_f \nu_f^{1/4}} \cdot \frac{d_f^{5/4}}{L} \cdot \Delta p \right)^{4/7} \tag{26}$$

when $Re > 2500$ (turbulent flow). Equation (25) is derived from the friction factor correlation for laminar flow in a parallel plate channel ($f = 96/Re$), and (26) from that for turbulent flow ($f = 0.316 Re^{-1/4}$) [16]. In (25) and (26) ρ_f is the density, ν_f the kinematic viscosity of the coolant. The maximum allowable coolant velocity (V_{fmax}) is concerned with corrosion prevention on the surface of the coolant channel, which is set around 2 m/s in many engineering applications. The coolant velocity is set at V_{fmax}, when V_f of (25) or (26) exceeds V_{fmax}.

The temperature rise of coolant is calculated from

$$\Delta T = \frac{Q_C}{\rho_f c_{pf} L d_f V_f} \tag{27}$$

The element temperature at the downstream end of the coolant channel (θ_S) is calculated using the heat transfer coefficient, $h = 4.115 k_f/d_f$ for laminar flow, and $h = 0.0395 Re^{0.755} Pr^{0.45} (k_f/2d_f)$ for turbulent flow [16], [17]. In these correlations, k_f is the thermal conductivity of coolant, and Pr is the Prandtl number.

$$\theta_S = \Delta T + \frac{q_C}{h}, \tag{28}$$

where $q_C = Q_C/L^2$ is the surface heat flux.

III. CALCULATION STEPS

The graph of computational density versus computational efficiency [7] provides a platform for our projection of the supercomputer technology in the past, the present, and the future. The computational efficiency is defined as

$$C_{eff} = \frac{S_{TP}}{Q_{sys}} \tag{29}$$

Its dimension is operations/J. The computational density is defined as, writing the dimension of the system volume in

liter (L),

$$C_{dns} = \frac{S_{TP}}{1000 \times V_{sys}} \tag{30}$$

where V_{sys} of (21) has the dimension of m^3. The dimension of C_{dns} is operations/(s· L).

All the model computers considered in the present study are supposed to be cooled by a dielectric coolant FC77, the same coolant employed in CRAY-1. The model computers include those that have equivalent existing machines and those still in the realm of speculation, as defined in Table 1. Model 'C' corresponds to CRAY-1 developed in 1976, 'E' to Earth Simulator of NEC (2002), 'K' to K-Computer of Fujitsu (2011), 'D' is a model to meet the target set by DARPA and the target year is 2015 [2]. Model 'Z' is a speculative computer capable of zeta-scale computing. Model 'B' is an equivalent of human brain in throughput, power consumption and volume. Table 1 shows the system data, where the throughput (S_{TP}), the system heat (Q_{sys}), and the system volume (V_{sys}) are those of the existing machines or the target values. Notes are due about some of the data. The throughput of supercomputers has been reported in the literature usually in FLOPS (floating point operations per second), while the measure of system performance in the present model is in the vein of bits per second (bps). In an energy-efficient fused-multiply-add unit about 50 bits in the data path participate in one FLOP [21]. However, in the C_{dns}-C_{eff} graph and others, which have the logarithmic axes spanning several orders of magnitudes, a factor of 50 has a minor influence on the location of the data plot. Hence, we make no distinction between FLOPS and bps. Estimation of the system volume of the existing machine involves a certain level of uncertainty. For example, K-computer is composed of 863 racks, but the rack contains a large fraction of extraneous space. We suppose that a PCB supporting CPU in one rack has an area 1 m^2, and each rack is 1 m deep; hence an essential volume participating computing is estimated as 1 m^3 per rack. The same rule of estimation is applied to the target volume of DRAPA project, which is specified only as 500 conventional server racks [2].

Figure 2 shows the graph of $C_{dns} - C_{eff}$ in which the data symbols represent the state points corresponding to S_{TP}, Q_{sys}, and $V_{sys}$of Table 1. The state points fall on the diagonal line which coincides with that in the graph produced by Ruch et al. [7]. The formulas developed in the preceding section involve a certain number of parameters. We use some of them as the knobs to bring the state point of a model computer fall on the points of their equivalent machines in Fig. 2. The rest of the parameters are fixed, common to different generations of model computers. The fixed parameters are listed in Table 2. The Rent exponent of $p = 0.8$ has been assumed in the modeling of high-performance computing [7]–[12]. The Rent constant A in the literature varies in a wide range, 1.4 [15, p. 453] to 10 [7, p. 15:9]; we adopt $A = 5$ [12].

The RC delay time (τ_C) is assumed to be independent of scaling [15]. Its value is set referring to several sources of interconnection dimensions and delay time data. On modern

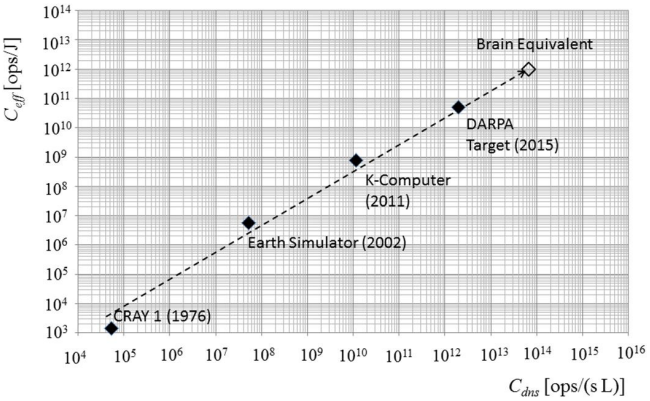

FIGURE 2. **Computational efficiency (C_{eff}) versus computational density (C_{dns}): The state points of the supercomputers of several generations are shown by solid diamonds. The state point of model B (Brain Equivalent) is shown by an open diamond.**

TABLE 1. **System data of model computers.**

Symbol	Equivalent computer	Year	S_{TP} [ops/s]	Q_{sys} [MW]	V_{sys} [m^3]	V_D [V]	Reference
C	CRAY-1	1976	1.6 x 10^8	0.115	3	2	[18, 19]
E	Earth Simulator	2002	3.6 x 10^{13}	6.4	700	1.5	[20]
K	K computer	2011	10^{16}	12.7	863	0.8	[1]
D	DARPA target	2015	10^{18}	20	500	0.5	[2]
Z	Zeta-scale computer		10^{21}			0.5	
B	Human brain		10^{14}	100 [W]	0.0015	0.1	[7]

TABLE 2. **Fixed parameters.**

- Physical constants:
  - Speed of light $c_0 = 3 \times 10^8$ [m/s]
  - Dielectric constant in vacuum $\varepsilon_0 = 8.854 \times 10^{-12}$ [F/m]
- Material properties:
  - Copper for power delivery $\rho_p = 1.72 \times 10^{-8}$ [Ωm]
  - Relative dielectric constant $\varepsilon_r = 2$
- Parameters in the Rent's rule:
  - Exponent $p = 0.8$
  - Constant $A = 5$
- Time overheads:
  - RC delay on the card $\tau_C = 0.76 \times 10^{-9}$ [s]
  - On the modulator/receiver for optical lines $\tau_{zd} = 2 \times 10^{-9}$ [s]
- Properties of coolant (FC77)
  - Specific heat at constant pressure $c_{pf} = 1172$ [J/kg K]
  - Thermal conductivity $k_f = 0.057$ [W/m K]
  - Prandtl number $Pr = 9.75$
  - Kinematic viscosity $\nu_f = 2.83 \times 10^{-7}$ [m^2/s]
  - Density $\rho_f = 1590$ [kg/m^3]

VLSI chips interconnection wires are laid in multiple layers, thick wires for global routing and thin wires for local interconnections. The scaling factor, (17), is estimated as $S_l = 4 \times 10^{-9}$ taking a median of the range of data; for example, the corresponding cross sectional area of wire corresponds approximately to that of wires of global tier at

the 50 nm technology node [22] and the wire length of 4 cm. Meanwhile, on a thinned copper wire the resistivity increases from that of bulk copper, almost by an order of magnitude [23]. When $\rho = 1.7 \times 10^{-7}$ [W· m] (10 × the value of bulk copper), $\varepsilon_r = 2$, and $S_l = 4 \times 10^{-9}$ are substituted in (16), we obtain $\tau_C = 0.76$ [ns]. This value coincides approximately with the upper bound of signal transmission time in CRAY-1 (∼1 ns, [18]). Thus, the above τ_C value represents a level of on-card delay time that covers the generations of supercomputers from CRAY-1 to the recent one employing thin on-chip copper wires. We must note that, as explained in the previous section, this τ_C value is used in the formula for power consumption (15), while the delay time which determines the system throughput is that for signal transmission across the card stack. Later in the calculation, the scaling parameter (S_l) will be used as one of the knobs to tune the state point towards the evolution line in the $C_{dns} - C_{eff}$ graph; hence, its value will have to be varied for computers of future generation. To keep the τ_C value at the stated level we need novel wiring materials which decrease the line resistivity in proportion to the variation of S_l, as required by (16). The value of relative dielectric constant, $\varepsilon_r = 2$, is chosen referring to the data of on-chip dielectric, which is 3.75 at the 180 nm technology node and projected as 1.25 at the 50 nm node [22]. This value is also close to the value of organic wiring substrate such as polyimide.

The time overhead at the modulator/receiver for optical lines at the card edges ($\tau_{zd} = 2$ ns) is chosen to be in a range of the rise time on on-silicon optical switches (1 – 8 ns [24]). Heat generated from the optical devices is assumed to be removed by a separate cooling design; hence heat from the optical components does not flow into the coolant in the channels between the cards. In essence, the system throughput is explicitly represented by the delay time on the modulator/receivers and the optical lines, while the cooling requirement is explicitly defined by the operation of on-card circuits. Correspondence between on-card and off-card (along the stack) operations is included in (11) in terms of the number of communication lines emerging at the card edges and the factor A_C.

A note is made about the motives to fix those parameters of Table 2. The actual values of the parameters vary from a generation of supercomputer to the next. But, their variations are deemed small in the analysis which involves several orders of magnitudes of variations in the primary variables and parameters. Fine tuning of all the parameters following the actual data is considered undue in the vein of present system modeling. An exception is made for the power bus voltage (V_D). The formula for the circuit element energy, (13), involves V_D^2; hence, its variation through the generation of equivalent supercomputers is taken into account as shown in Table1.

There are a number of parameters other than those mentioned above. Counting those directly relevant to our analysis we have 30 parameters plus C_{dns} and C_{eff}; hence 32 parameters are to be determined or calculated. We count the number of equations in the preceding section, excluding those used for intermediate explanations and including those unnumbered ones such as $\bar{l}_C = \bar{R}_C d_e$ and others, as 26. When (29) and (30) are included, the number of equations amounts to 28. Hence, 4 parameters are free to choose. Several ways to specify these parameters and perform calculations are conceivable. The charts in Fig. 3 show several processes of calculation.

For the models of Table 1 except for model Z, S_{TP}, Q_{sys}, and V_{sys} are specified. Hence, for these models the number of free parameters is 1, which we identify with either A_C or γ_A. The A_C is the activity factor for longitudinal (stack-wise) communications, and γ_A is the fraction of active on-card circuits at any instant. The process labeled 'Start' in Fig. 3 is for the initial reference computer, which is CRAY-1. We set $\gamma_A = 1$; we suppose that the element circuit in the reference computer is actually a set of finer level circuits and any components of the element circuit is always active. The first-phase calculation is performed for the system heat equation, (19), into which the relevant parameters starting with P_{eH} from (18) and those from (13)–(15) are substituted. In this process 11 equations are used for 11 undecided parameters, and (19) is reduced to the equation for the system size N_S. The Newton-Raphson iteration is applied to the resultant equation to determine N_S. The second-phase calculation is to determine A_C from the system throughput equation, (11). The parameters other than A_C involved in (11) are either specified (S_{TP}) or calculated using the determined value of N_S. All the parameters including the card thickness components (d_{sp}, (22)), d_{ss} ((23)), the coolant path width (d_f, from (24)), the coolant

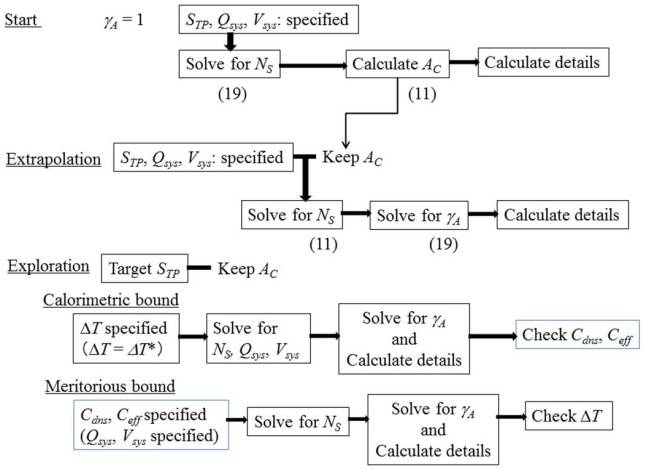

FIGURE 3. Flow of calculations: 'Start' is to determine A_C, the relative measure of across-the-stack communication activity, which is assumed invariant throughout the computer generations. 'Extrapolation' is applied where the computational density (C_{dns}) and the computational efficiency (C_{eff}) are calculable, while the details about the hardware of the stacked-card model need the value of γ_A, the relative measure of on-card activity. In 'Exploration' only the target value of the system throughput (S_{TP}) is specified, and two extreme conditions, 'Calorimetric bound' and 'Meritorious bound', are considered. In 'Calorimetric bound' the coolant temperature rise between the inlet and the exit is set at the threshold value (ΔT^*). In 'Meritorious bound' the figure-of-merit pair (C_{dns}, C_{eff}) is held unchanged, which specifies the system volume (V_{sys}) and the power consumption (Q_{sys}) to realize a target S_{TP}.

velocity (V_f, (25) or (26)) and the temperature rise (ΔT, (27)), and the maximum surface temperature of the card (θ_S, (28)) are calculated in the phase labeled as 'Calculate details' in Fig. 3.

The process labeled as 'Extrapolation' in Fig. 3 is applied to the later generations of model supercomputers, from E to D, and the model equivalent to human brain, B, where S_{TP}, Q_{sys}, and V_{sys} are specified. In this process we keep the value of A_C determined in the previous process 'Start', while we treat γ_A as an undetermined parameter. This exchange of value assignment needs a note. In (11), the product $M \cdot N_C^p$ on the right hand side is the total number of signal lines emerging from the card edges. The A_C is the system design factor which reflects the ratio of on-card processing to off-card communications. This ratio is supposed to be a slowly varying variable through computer generations; hence, to the first order approximation, we use the value determined for the initial reference machine for later generations. Meanwhile, the circuits on the card undergo evolutions of considerable degrees from one generation to the next. Miniaturization of devices and wirings, and co-implementation of different functional elements, is the notable evolutionary features we observe in the chips and the printed circuit boards of actual computers. The circuit element in the present model has to reflect such evolutions. We incorporate the evolutionary feature by reducing γ_A, supposing that circuit elements embody different functional blocks, hence, the element is only intermittently active. In the first-phase of calculation we work with the system throughput equation, (11). For a given set of S_{TP} and V_{sys}, (11) is transformed to the equation to determine N_S. The Newton-Raphson iteration is applied to solve the equation for N_S. For a determined N_S all the parameters but γ_A are calculated, and the system heat equation, (19), is reduced to the equation for γ_A. Again, the Newton-Raphson iteration is applied, this time, to determine γ_A. Calculations of the geometrical and thermal details follow in the phase 'Calculate details' as in the process 'Start'.

The process 'Exploration' in Fig. 3 is for future generation computers where we have a target value for the system throughput (S_{TP}) but no specific data for Q_{sys} and V_{sys}. We explore the boundaries of the parametric domain defined by two constraints, 'calorimetric' and 'meritorious'. The calculation process labeled as 'Calorimetric bound' in Fig. 3 is for a scenario where reduction of the system volume (V_{sys}) is aggressively pursued while the system heat (Q_{sys}) is adjusted to keep the state point on the evolutionary line in the $C_{dns} - C_{eff}$ graph. The limit to volume reduction is set by the temperature rise of coolant in an extremely narrowed coolant path. We set $\Delta T = \Delta T^*$ in (27), where ΔT^* is the maximum allowable temperature rise of the coolant. (Another thermal criterion regarding the surface temperature (θ_S) does not play a critical role, as we will see in the numerical examples in the next section.) To let the state point stay on the evolutionary line we introduce the following equation.

$$Q_{sys} = Q'_{sys} + \frac{1000}{\Delta C}\left(V_{sys} - V'_{sys}\right) \tag{31}$$

where Q'_{sys} and V'_{sys} belong to a computer of the current generation, hence, known. The ΔC is the slope of the evolutionary line in the $C_{dns} - C_{eff}$ graph (Fig. 2), set as 0.0677. The calculation involves 18 parameters and 18 equations including (31), and after some manipulations the equations are reduced to the one to determine N_S. In this process the system heat equation, (19), is set aside, because (31) serves as its complement. After determination of N_S, (19) is invoked to determine γ_A for the calculation of the details. After all the parameters are determined, the values of C_{dns} and C_{eff} are calculated to find the state point in the evolution graph. By this calculation we find possible improvement in the measures of C_{dns} and C_{eff} for a system where ΔT is allowed to rise to a threshold value (ΔT^*). It may happen that one or both of the calculated values of C_{dns} and C_{eff} are below the corresponding value(s) of the current generation computer. Such a result means that the improvement on the evolutionary line is impossible in a system of aggressively squeezed volume, unless some novel materials are introduced in the future generation computer. The relevant parameter for line material is the scaling factor, S_l, which will be varied in the case studies of the next section to explore ways to overcome the constraint.

The 'Meritorious bound' sets a relaxed scenario where Q_{sys} and V_{sys} are allowed to increase, but the values of C_{dns} and C_{eff} are kept equal to those of a current generation computer; hence, the meritorious indexes are not allowed to deteriorate in the next generation computer. Since, C_{dns} and C_{eff} are fixed, Q_{sys} and V_{sys} are determined directly from the definitions of these indexes ((29) and (30)) for a target S_{TP} of the next generation computer. The rest of the calculation process is the same with that of 'Extrapolation'. At the end of the calculation we check if ΔT is below the threshold ΔT^*. It may happen that the resultant ΔT exceeds ΔT^*. Such a result means that, even following a relaxed scenario, the target S_{TP} cannot be attained, unless some novel materials are introduced in the future generation computer. Again, the scaling factor S_l will be varied in the case studies of the next section to explore ways to overcome the constraint.

As already mentioned, the 'Start' process is applied to model C, the equivalent of CRAY-1. The process 'Extrapolation' is performed on models E, K, D, and B. It will be shown that the realization of D, the equivalent of DARPA target computer, may need novel materials for on-line signal lines. The process 'Exploration' is used to explore the possibility of zeta-scale computer. It is also used to explore the evolutionary step from model K to variants of model D.

IV. RESULTS AND DISCUSSION

Case studies were conducted assigning values to some parameters in addition to those included in Tables 1 and 2. For C, E, K, and D, we set $L = 1$ m, and for B, $L = 0.1$ m. The scaling factor S_l is set at 4×10^{-9} for C, E, K, and B; but a lower value is needed for D. The scaling factor is further treated as a variable in the calculation process 'Exploration'. The factor to account for Joule heating on the power/ground lines is estimated as $f_{JH} = 1.18$ (in (18)), and the factor to determine

the voltage drop allowance is set as $\alpha_R = 1.42$ (in (22)). (See Appendix about these factors.) The figures concerning the fluid-thermal criteria are set as follows; the pressure drop in the coolant path $\Delta p = 20$ kPa; the maximum allowable coolant velocity $V_{fmax} = 2$ m/s; the maximum allowable temperature rise of coolant $\Delta T^* = 50$ K, which gives 80° for the coolant of 30° at the inlet. From the process 'Start' we derived $A_C = 3.45 \times 10^{-6}$, which is maintained in all calculations. The calculated results and discussions will be made in the following sub-sections; section *A* reports the results for models C, E, K, D, and B, where the process 'Extrapolation' is applied; in section *B* our primary concern is the challenges of designing a zeta-scale computer.

A. ON MODELS C, E, K, D, AND B

Figure 4 shows the plot of system throughput (S_{TP}) versus calculated system size (N_S) for models C, E, K, D, and B. For the exa-scale computer, D, the required system size is $N_S = 1.38 \times 10^{18}$, about 100 times that for K (1.51×10^{16}) which yields 10 x peta-scale throughput. Hence, from K to D, the number of required circuit elements increases almost in proportion to the increase of system throughput. The partition policy (9) holds the increase of cards in the stack by a factor of about 5; from $M = 1.93 \times 10^5$ for K to 1.08×10^6 for D. To accommodate the required increase of N_S, the population of circuits on the card must increase by a factor of 13, from $N_C = 2.66 \times 10^{11}$ for K to 3.49×10^{12} for D. This requires reduction of the circuit element's area by an inverse proportion; in terms of the side length of the circuit element, from $d_e = 1.94$ μm for K to 535 nm for D. Although miniaturization of circuit elements on the cards bears by a large proportion the impact of system size increase, the number of cards in the stack (M) grows to the level of a million in the evolution from K to D. To contain the system volume of D the card thickness and the coolant path width must be reduced to very small dimensions, as we will see shortly. Also for D, the scaling factor S_l has to be reduced by a factor of 10, that is, $S_l = 4 \times 10^{-10}$. Otherwise, the specifications of S_{TP}, Q_{sys}, and V_{sys} for D (Table 1) cannot be met. (About the relationship between S_l and the figures of merit such as C_{dns} and C_{eff} will be discussed later in this section.) To keep the time τ_C of (16) unaltered despite the reduction of S_l we need to suppose that the resistivity of signal line (ρ) is lowered by introduction of a new material for signal transmission lines. Reduction of ρ by an order of magnitude from that of copper thin line can be made possible by introduction of carbon nanotubes [23].

The model of equivalent human brain, B, is capable of yielding the throughput of a supercomputer of almost 100 times larger in the system size (N_S), as observed in Fig. 4. The point closest to B in Fig. 4 is E where $S_{TP} = 3.6 \times 10^{13}$ and $N_S = 1.16 \times 10^{14}$, while at B $S_{TP} = 10^{14}$ and $N_S = 7.84 \times 10^{12}$. Such performance of B owes primarily to the small volume of B which cuts the signal transmission time by more than an order of magnitude from that in a computer yielding comparable throughput. For example, the time τ_z of (12) is 60 ns in E, while it is reduced to 2 ns in B. Tight spatial constraint in B makes the allowance for coolant passage extremely small as shown in Fig. 5.

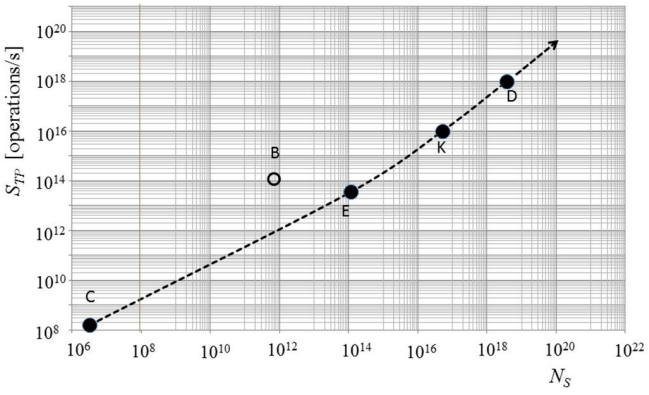

FIGURE 4. Correspondence between the system throughput (S_{TP}) and the total number of circuit elements in model computers (N_S).

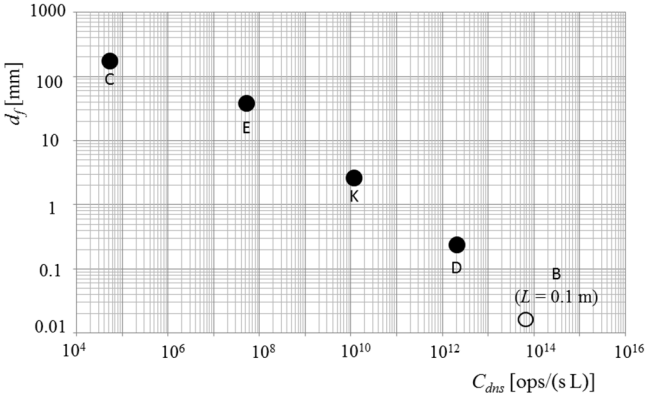

FIGURE 5. Coolant path width (d_f) versus computational density (C_{dns}).

Fig. 5 shows the coolant path width (d_f) versus the computational density (C_{dns}). For those equivalent models of the existing computers, C, E, and K, certain values are assumed for d_{s0} in (24), 1 cm for C, 2 mm for E, and 1 mm for K. These values are substituted in (24) to calculate d_f. However, the assumption about d_{s0} is only to mimic actual hardware; hence, it introduces no substantial alteration to what we learn from Fig. 5. An issue of our primary interest is how tight the spatial allowance for coolant flow paths could be in the future generation computers including the equivalent of human brain. To estimate the maximum allowance for d_f we set $d_{s0} = 0$ in the calculations for models D and B. We find from Fig. 5 that the coolant path width has to decrease by an order of magnitude from one generation to the next, and in D the maximum allowance for d_f shrinks to 240 μm, and in B to 16 μm. The coolant passages having these dimensions fall in the category of micro-channels. Heat transfer in micro-channels has been a subject of active heat transfer research since the pioneering work of Tuckermann and Pease [25]. However, almost all of the works on micro-channel heat transfer have had their focus on chip-level cooling; that is, the coolant path length is in a range 1–2 cm. The length to

width ratio of coolant path is around 200 in the case of chip cooling; by contrast, the ratio in the prospective computers is about 4000 in D and 6000 in B. We note here that the dimensions of coolant path in chip-cooling result from optimized balance between calorimetric resistance and surface heat transfer, and the objective of optimization is to attain lowest thermal resistance on the chip [25]. The objective to achieve lowest possible thermal resistance has been inherited in those subsequent studies after [25], including the work on chip-stack cooling [3]–[5]. By contrast, the micro-scale dimensions of coolant paths in the prospective computers result from the demand for extreme compact packaging of the system. Meanwhile, the heat load for each coolant path is not as high as those assumed in the chip-cooling studies, as we will see next.

The heat load per card (Q_C) and the surface heat flux on the card (q_C) are plotted against the computational density (C_{dns}) in Fig. 6. These data show remarkably low demands on cooling design. The often-cited heat load of 100 W on a high-performance chip corresponds to a heat flux level of 10^6 W/m^2, out of the scale of the vertical axis. The coolant velocity is constrained by the upper bound of 2 m/s in models C, E, and K; while in D and B, the velocity is the result of applied pressure difference between the inlet and outlet of the coolant path (20 kPa) and the flow is laminar (0.21 m/s in D, 1 cm/s in B). The resultant temperature rise in the coolant is quite small; 0.2 K in D and 0.6 K in B. The surface temperature at the downstream end of the coolant (θ_S of (28)) is above ΔT due to surface thermal resistance; in the models of existing computers, $\theta_S = 3$ K versus $\Delta T = 0.01$ K in C, 0.11 K versus 0.003 K in E, and 0.016 K versus 0.007 K in K; however, θ_S is negligible in D and B due to high heat transfer coefficients in the reduced-size channels. Low level heat loads, particularly in the prospective computers D and B, are the result of decrease in γ_A, the factor representing the activity of the circuit element, and γ_A has to be decreased to have the state point stay on the evolution line of the C_{dns}-C_{eff} graph. To check the reasonableness of the present model we calculated the energy required to drive a line of unit length, E_e/l_e ((13), (14)). The figures, 35 x 10^{-12} J/m in C and 6 $\times$ 10^{-12} J/m in K, are comparable to 39 $\times$ 10^{-12} J/m at the 130 nm technology node and 7 $\times$ 10^{-12} J/m at the 45 nm node (these figures at the technology nodes from [7, p. 15:3]).

B. ON MODELS D AND Z

The calculation process 'Exploration' is applied to explore the possibility of exa-scale computing and beyond. The parameter of primary importance in this exploration is the scaling factor S_l, (17), which defines the cross sectional area (A_l) of a signal line of average length on the card ($\bar{l}_C$). The required thickness to accommodate signal lines at the card edge, d_{ss}, is given by (23). Although A_l is less than the cross sectional area of the power-ground line by three orders of magnitude (not detailed to contain the paper length), the number of signal lines at the card edge increases with increasing N_C (the number of circuits elements on the card), and the required space

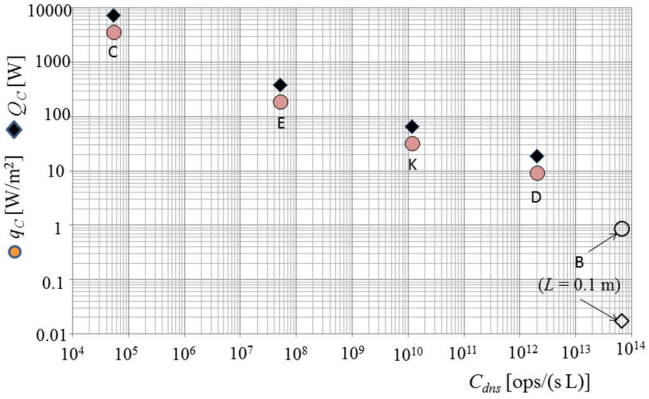

FIGURE 6. Heat dissipation per card (Q_C) and surface heat flux (q_C) versus computational density (C_{dns}), calculated for the models.

for signal lines at the card edge becomes a dominant component in the prospective computers. Where the component d_{ss} reaches the card pitch, d_p, which is defined by the system volume (21), there remains no space for coolant, hence, no possibility to realize a prospective computer, unless A_l is decreased through modification of S_l. In changing $S_l$the delay time τ_C (16) is made invariant; imposition of this condition is tantamount to the effort to keep the on-card delay time from increasing with thinning signal lines. We assume that S_l can be decreased by employing materials of low resistivity for signal transmission lines.

Figs. 7 and 8 show the roles of the scale factor S_l in our attempt to upgrade the performance of K, a peta-scale computer, to exa-scale. Fig. 7 shows the results of the calculation process 'Calorimetric bound' in the $C_{dns} - C_{eff}$ graph; the solid diamond symbols show the state points belonging to exa-scale computers ($S_{TP} = 10^{18}$) of different scale factors, and an open circular symbol shows the point of K. At all the sate points of the variants of D the coolant's temperature rise is 50K. As we reduce the scale factor from that of K, i.e. $S_l = 4 \times 10^{-9}$, we obtain improvements in the figures of merit of C_{dns} and C_{eff}. The improvements are brought through beneficial interactions of several parameters, among which the reduction of card pitch (d_p) has direct impact on the system volume and the system power. To achieve a specified exa-scale throughput ($S_{TP} = 10^{18}$) the requirement for system size (N_S) decreases with decreasing d_p. This can be understood by examination of (11) as follows. The stack length decreases with decreasing d_p, and a reduced stack length decreases the transmission time along the stack, τ_z. Since S_{TP} on the left hand side of (11) is given, the reduction of τ_z is matched by the reduction of $M \cdot N_C^p$ on the right hand side, which in turn means reduction of the required N_S. A system with smaller N_S requires less power and volume; hence, the improvement in the efficiency and density results.

The point for an exa-scale computer with $S_l = 4 \times 10^{-10}$ (in the upper right corner) has much higher figures of merit than model D for which the equal value of S_l is assumed.

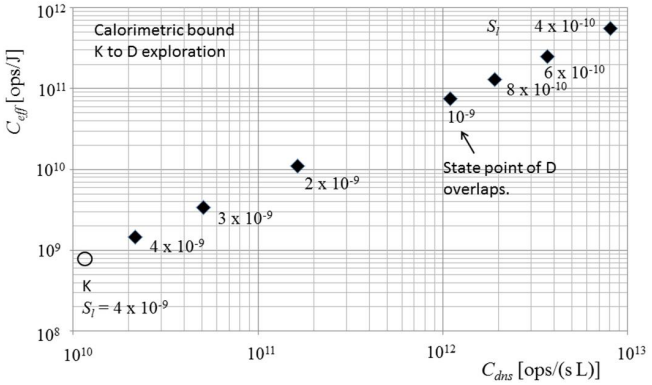

FIGURE 7. Shift of the state point of exa-scale variants of D (solid symbols) effected by the change of scaling factor S_l. The D variants are on the calorimetric bound (ΔT = 50K). The open symbol is the point of peta-scale model K.

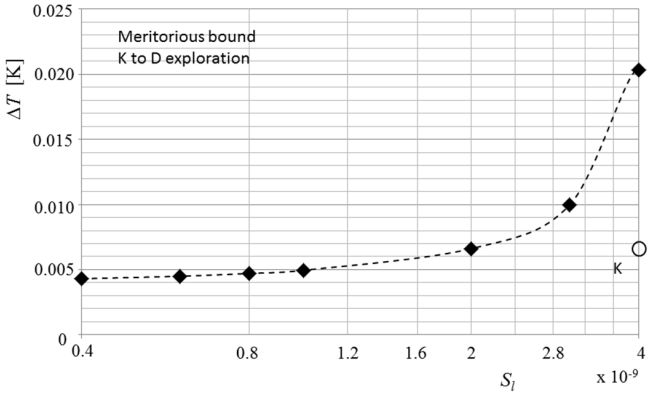

FIGURE 8. Coolant temperature rise (ΔT) versus scaling factor (S_l); the solid symbols are for the variants of exa-scale model D, and the open symbol for peta-scale model K. Models variant-D and K in this graph have equal figures of merits (C_{dns} = 1.16 × 10^{10}, C_{eff} = 7.87 × 10^{8}).

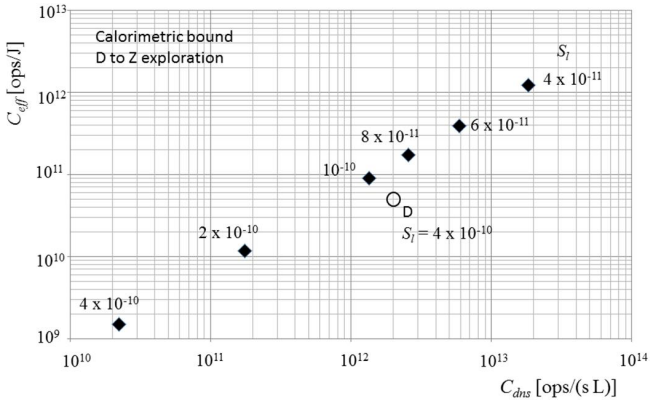

FIGURE 9. Shift of the state point of zeta-scale model Z (solid symbols) effected by the change of scaling factor S_l. The variants of Z are on the calorimetric bound (ΔT = 50 K). The open symbol is the point of exa-scale model D.

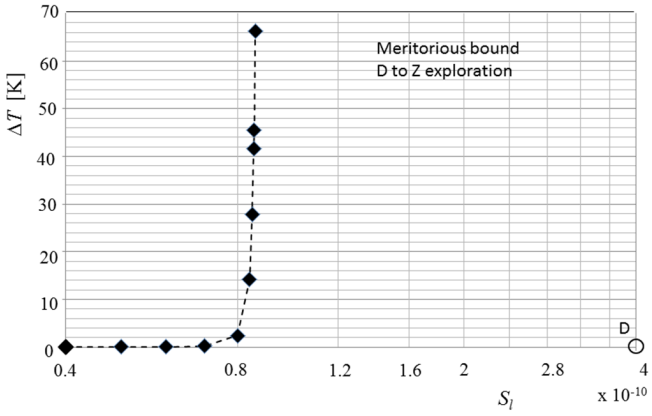

FIGURE 10. Coolant temperature rise (ΔT) versus scaling factor (S_l); the solid symbols are for zeta-scale model Z, and the open symbol for exa-scale model D. Models Z and D in this graph have equal figures of merits (C_{dns} = 2.00 × 10^{12}, C_{eff} = 5.00 × 10^{10}).

The state point of D overlaps the symbol for $S_l = 10^{-9}$ in Fig. 7. This set of relatively low figures of merit on D is the consequence of small ΔT (0.2 K) associated with the system volume specified as the DARPA target. If the system volume is squeezed to a level where ΔT approaches 50K, the figures of merit would improve by about an order of magnitude. Meanwhile, the point of $S_l = 4 \times 10^{-9}$ in Fig. 7 indicates that, even where new materials are not employed to decrease S_l, but ΔT is allowed to approach 50K, exa-scale throughput is still possible with modest improvements in C_{dns} and C_{eff}.

Fig. 8 shows the results of the calculation process 'Meritorious bound', ΔT versus S_l. The solid diamond symbols belong to exa-scale computers, while the open circle to model K. Decrease of S_l on the exa-scale class of computers yields reduction of ΔT. The ΔT in an exa-scale computer employing the same S_l assumed in K increases by four times that in K, but is still at a low level. In the scenario of 'Meritorious bound', however, the system heat and volume must increase by 100 times those of K. This is obvious from the definitions of C_{eff} (29) and C_{dns} (30).

The results shown in Figs.7 and 8 imply that the evolutionary step from the peta-scale to the exa-scale computer is challenging but still achievable without serious deterioration in the figures of merit or violation of the temperature criterion regarding ΔT. The step towards zeta-scale computing ($S_{TP} = 10^{21}$), however, requires some disruptive developments of materials for signal transmission and heat transport. Fig. 9 shows the plot of state points for zeta-scale computers (solid diamonds) and the one for model D (open circle). The numerical figure attached to the data symbol is the value of S_l. If we keep the same S_l used in D (4×10^{-10}), C_{dns} decreases by two orders of magnitude, and C_{eff} by almost one order of magnitude. To attain the equal levels of density and efficiency with those of D S_l has to be decreased to fourth or fifth of the value assumed for D. Fig. 10 is the graph of ΔT versus S_l, where the solid diamonds are the data points belonging to zeta-scale computers and the open circle to model D. Although the system volume and heat are allowed to grow by three orders of magnitude in the scenario of 'Meritorious bound', ΔT in zeta-scale computers shoots up to prohibitively high level, unless S_l is decreased below a certain threshold. The threshold S_l is close to 8×10^{-11} in Fig. 10. Rapid increase of ΔT near the threshold reflects the loss of space for coolant due to the increasing requirement for N_S with increasing S_l. An alternative way to

push back the constraint posed by the increase of ΔT is the use of a coolant which has a larger heat capacity than that of FC77. Water provides the heat capacity 2.2 times that of FC77. Phase change of the coolant absorbs heat in the form of latent heat, hence, suppresses the rise of ΔT. Endothermic chemical reaction in fluid, used in heat storage applications, offers another measure to suppress ΔT. These candidates for high-capacity coolant require considerable research and development efforts to be used in long micro-channels that run through the blocks of microelectronic circuits.

V. CONCLUSION

In the present study the hardware construction of a computer is modeled by a stack of integrated circuit cards. This simplified configuration allows us to develop a set of equations that describe the relationships between various parameters. The parameters of interest are the processing throughput in operations per second, the system volume, the system power, and those parameters concerning the details of internal organization such as the signal and power line dimensions and the coolant flow path width. The objective of the study is to elucidate challenges facing the hardware design of future generation computers. Towards this objective we employ the graph of a pair of the figures of merit, the computational density and the computational efficiency. The evolution of supercomputers of the past and the present generations is captured by a string of state points in the density-efficiency graph. A few adjustable parameters embedded in the model are used as the knobs to steer the calculation results along the evolution trend formed by the state points of the existing machines. Thus developed calculation steps are employed to locate the state points of future prospective computers on the evolution curve, including an equivalent of human brain. The details of internal organization are then calculated to find the impacts of extremely tight spatial and power constraints on them.

The calculation results for prospective exa-scale and zeta-scale computers point to the needs of research and development, some of which are off the focus of current research efforts. The most notable departure is found in the needs for cooling design. In the thermal management community today one of the popular research topics is how to deal with possible high heat flux in chip stacks. However, the heat flux could not be high in future very-large-scale computers due to tight spatial and power budgets. Instead, low heat in very long microfluidic channels could be the norm. Low heat per coolant channel, however, does not mean the dissolution of thermal challenges. Where the coolant flow is slowed below a certain threshold by accidental cause, a thermal crisis will appear abruptly. Hence, the microfluidic robustness is the key to the design of cooling systems. Further towards the realization of zeta-scale computing we need novel materials that require far less space than copper lines. Another barrier for zeta-scale computing is the temperature rise of coolant in extremely narrowed coolant paths. To overcome the calorimetric bound, a coolant having large apparent heat capacity will be required.

APPENDIX

Power is supplied through the power terminals on the card edges, and each power terminal supplies electric current to a row of circuit elements. We suppose power and ground lines embedded in the card, and individual circuit elements are smeared out in a continuous model. The power consumption per unit length of the homogenized row is P_e/d_e, where $P_e$is the power consumption by the circuit element of length d_e. The voltage on the power line (V_p) decreases from the edge into the depth of the row (x), and that on the ground line (V_g) increases along x. Denoting $\Delta V = V_p - V_g$we write the local current balance as

$$\frac{dI}{dx} = -\frac{P_e}{d_e \cdot \Delta V} \tag{A1}$$

The voltage drop along the power line is written as

$$V_p = V_D - R_p \int_0^x I \cdot dx \tag{A2}$$

where R_p is the electrical resistance per unit length of the power line, and calculated from $R_p = \rho_p / A_{pe}$, where ρ_p is the resistivity and A_{pe} is the cross sectional area of the power line. At $x = 0$, $V_p = V_D$ (supply voltage). The ground line voltage (V_g) is written likewise, but with $V_g = 0$ at $x = 0$. Substituting V_p and V_g into ΔV we rewrite (A1) as

$$\left(V_D - 2R_p \int_0^x I \cdot dx \right) \cdot \frac{dI}{dx} = -\frac{P_e}{d_e} \tag{A3}$$

Equation (A3) is further integrated from $x = 0$ to $x = L/2$. At $x = L/2$ (half width of the card), $I = 0$ because of the symmetry. Using this condition, the integral form of (A3) is now written as

$$V_D I_r - 2R_p \int_0^{L/2} I^2 \cdot dx = \frac{P_e \cdot L}{2d_e} \tag{A4}$$

From (A3) we have a condition at $x = 0$: $dI/dx = -P_e/V_D d_e$. The form of I satisfying the conditions at $x = 0$ and $L/2$ is written as

$$I = I_r - \frac{P_e}{V_D d_e} x - \frac{4}{L^2} \left(I_r - \frac{P_e L}{2 V_D d_e} \right) x^2 \tag{A5}$$

where $I_r$is the electric current at the power terminal. Substituting (A5) into (A4), we have a quadratic equation for I_r. The solution is written as

$$I_r = \frac{1}{2} \left[B - \sqrt{B^2 - 4C} \right] \tag{A6}$$

where

$$B = \frac{15}{8} \left(\frac{V_D}{R_p L} - \frac{7}{60} \frac{P_e L}{V_D d_e} \right) \tag{A7}$$

$$C = \frac{1}{64} \left(\frac{P_e L}{V_D d_e} \right)^2 + \frac{15}{16} \cdot \frac{P_e}{R_p d_e} \tag{A8}$$

The condition $B^2 > 4C$ sets the upper bound for R_p as

$$R_p* = \frac{60}{39+16\sqrt{6}}\left(\frac{V_D^2 d_e}{P_e L^2}\right) \tag{A9}$$

Meanwhile, Joule heating on the power/ground lines per element (averaged over $L/2$) is written as

$$P_{eJ} = 2\frac{R_p d_e}{L/2}\int_0^{L/2} I^2 \cdot dx \tag{A10}$$

Substituting (A5) into (A10) and writing the heat dissipation per element (element thermal power) as $P_{eH} = P_e + P_{eJ}$, we have

$$P_{eH} = P_e + \frac{2R_p * d_e}{15\alpha_R}\left\{8I_r^2 - \frac{7}{4}I_r\left(\frac{P_e L}{V_D d_e}\right) + \frac{1}{120}\left(\frac{P_e L}{V_D d_e}\right)^2\right\} \tag{A11}$$

where we set $R_p = R_p*/\alpha_R$, using a factor α_R (>1). Substituting further (A6) into (A11) and using (A9) we obtain

$$f_{JH} \equiv \frac{P_{eH}}{P_e} = 1 + \frac{1}{32a'\alpha_R}\left[\begin{array}{l}(a'\alpha_R)^2 - 53a'\alpha_R + (1022/15) \\ +(14 - a'\alpha_R)\sqrt{(a'\alpha_R)^2 - 78a'\alpha_R - 15}\end{array}\right] \tag{A12}$$

where $a' = 39 + 16\sqrt{6} = 78.192$.

The value of α_R is chosen so that the voltage drop at the middle of the card ($(V_p)_x = L/2$) is held below a specified tolerance. Here, we specify that 10% of V_D is a tolerable voltage drop at the middle of the card. For this condition we find $\alpha_R = 1.42$, and $f_{JH} = 1.182$.

REFERENCES

[1] N. Leavitt, "Big iron moves toward exascale computing," *IEEE Comput.*, vol. 45, no. 11, pp. 14–17, Nov. 2012.
[2] P. Kogge, "The tops in FLOPS," *IEEE Spectrum*, vol. 48, no. 2, pp. 44–50, Feb. 2011.
[3] J.-M. Koo, S. Im, L. Jiang, and K. E. Goodson, "Integrated microchannel cooling for three-dimensional electronic circuit architectures," *ASME J. Heat Transf.*, vol. 127, no. 1, pp. 49–58, Jan. 2005.
[4] T. Brunschwiler, S. Paredes, U. Drechsler, B. Michel, W. Cesar, Y. Leblebici, B. Wunderle, and H. Reichl, "Heat-removal performance scaling of interlayer cooled chip stacks," in *Proc. 12th IEEE ITHERM*, Las Vegas, NV, USA, Jun. 2010, pp. 1–12.
[5] B. Dang, M. S. Bakir, D. C. Sekar, C. R. King, and D. Meindl, "Integrated microfluidic cooling and interconnects for 2-D and 3-D chips," *IEEE Trans. Adv. Packag.*, vol. 33, no. 1, pp. 79–87, Feb. 2010.
[6] P. G. Emma and E. Kursun, "Is 3D chip technology the next growth engine for performance improvement?" *IBM J. Res. Develop.*, vol. 52, no. 6, pp. 541–552, Nov. 2008.
[7] P. Ruch, T. Brunschwiler, W. Escher, S. Paredes, and B. Michel, "Toward five-dimensional scaling: How density improves efficiency in future computers," *IBM J. Res. Develop.*, vol. 55, no. 5, pp. 15:1–15:13, Sep./Oct. 2011.
[8] J. Dongarra, "Trends in high-performance computing," *IEEE Circuits Devices Mag.*, vol. 22, no. 1, pp. 22–27, Jan./Feb. 2006.
[9] P. W. Coteus, J. U. Knickerbocker, C. H. Lam, and Y. A. Vlasov, "Technologies for exascale systems," *IBM J. Res. Develop.*, vol. 55, no. 5, pp. 14:1–14:12, Sep./Oct. 2011.
[10] W. E. Donath, "Placement and average interconnection lengths of computer logic," *IEEE Trans. Circuits Syst.*, vol. 26, no. 4, pp. 272–277, Apr. 1979.
[11] W. Nakayama, "On the accommodation of coolant flow paths in high-density packaging," *IEEE Trans. Compon., Hybrids, Manuf. Technol.*, vol. 13, no. 4, pp. 1040–1049, Dec. 1990.
[12] H. M. Ozaktas and J. W. Goodman, "Implications of interconnection theory for optical digital computing," *Appl. Opt.*, vol. 31, no. 26, pp. 5559–5567, Sep. 1992.
[13] D. Stroobandt and J. V. Campenhout, "Estimating interconnection lengths in three-dimensional computer systems," *IEICE Trans. Inf. Syst.*, vol. E80-D, no. 10, pp. 1024–1031, Oct. 1997.
[14] A. Rahman and R. Reif, "System-level performance of three-dimensional integrated circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 8, no. 6, pp. 671–678, Dec. 2000.
[15] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Reading, MA, USA: Addison-Wesley, 1990.
[16] F. P. Incropera and D. P. DeWitt, *Fundamentals of Heat and Mass Transfer*, 4th ed., New York, NY, USA: Wiley, 1996.
[17] W. M. Kays and M. E. Crawford, *Convective Heat and Mass Transfer*, 2nd ed., New York, NY, USA: McGraw-Hill, 1980.
[18] R. D. Levine, "Supercomputers," *Sci. Amer.*, vol. 46, no. 1, pp. 118-135, Jul. 1982.
[19] R. M. Russell, "The CRAY-1 computer system," *Commun. ACM*, vol. 21, no. 1, pp. 63–72, Jan. 1978.
[20] (2013). *Earth Simulator* [Online]. Available: http://en.wikipedia.org/wiki/Earth_Simulator
[21] S. Galal and M. Horowitz, "Energy-efficient floating-point unit design," *IEEE Trans. Comput.*, vol. 60, no. 7, pp. 913–922, Jul. 2011.
[22] K. Banerjee and A. Mehrotra, "Global interconnect warming," *IEEE Circuits Devices*, vol. 17, no. 5, pp. 16–32, Sep. 2001.
[23] A. Ceyhan and A. Naeemi, "Cu interconnect limitations and opportunities for SWNT interconnects at the end of the roadmap," *IEEE Trans. Electron Devices*, vol. 60, no. 1, pp. 374–382, Jan. 2013.
[24] A. V. Rylyakov, C. L. Schow, B. G. Lee, W. M. J. Green, S. Assefa, F. E. Doany, M. Yang, J. V. Campenhout, C. V. Jahnes, J. A. Kash, and Y. Vlasov, "Silicon photonic switches hybrid-integrated with CMOS drivers," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 345–354, Jan. 2012.
[25] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sink for VLSI," *IEEE Electron Device Lett.*, vol. 2, no. 5, pp. 126–129, May 1981.

WATARU NAKAYAMA (M'88–SM'90–F'03–LF'13) has been working on thermal management of electronic equipment. During his twenty-strong years of association with Hitachi's Mechanical Engineering Research Laboratory, he played pivotal roles in developing thermal solutions for the company's diverse electronic products. In 1990, he joined the Faculty Member with the Tokyo Institute of Technology (Tokyo Tech), Tokyo, Japan, where he taught and conducted research on microelectronic packaging with a focus on power and thermal management of computers. He served as a Visiting Professor with the University of Maryland, College Park, MD, USA, from 1996 to 2001. He is currently an International Research Consultant advising on the thermal management research projects in Japan.

He has received many prominent awards, including the ASME Heat Transfer Memorial Award in 1992, the ICHMT Fellowship Award in 1996, the ASME Electrical and Electronic Packaging Division Award (now Allan Kraus Memorial Award) in 1996, the JSME Award for Longstanding Contributions to Mechanical Engineering in 1997, the ITHERM Achievement Award in 2000, the InterPack Achievement Award in 2001, the Thermi Award in 2006, the JSME Funai Special Award in 2007, and the ASME Max Jakob Memorial Award in 2013. He is a fellow of ASME, a Life Member of JSME, and an Honorary Member of the Heat Transfer Society of Japan.