Frontiers in Heat and Mass Transfer (FHMT), 15, 12 (2020)
DOI: 10.5098/hmt.15.12

Global Digital Central
ISSN: 2151-8629

Frontiers in Heat and Mass Transfer

Available at www.ThermalFluidsCentral.org

THERMAL MANAGEMENT OF DATA CENTERS UNDER STEADY AND TRANSIENT CONDITIONS

Yogesh Jaluria[a,*], Arvindh Sunder[a], Jingru Z. Benner[b]

[a] *Mechanical Engineering Department, Rutgers University, Piscataway, NJ 08854, USA*
[b] *Department of Mechanical Engineering, Western New England University, Springfield, MA 01119, USA*

ABSTRACT

Data centers are of crucial importance today in the storage and retrieval of large amounts of data. Most organizations and firms, ranging from banks and online retailers to government departments and internet companies, use data centers to store information that can be recovered efficiently and rapidly. As the deployment of data centers has increased, along with their capacity for data storage, the demands on thermal management have also increased. It is necessary to remove the energy dissipated by the electronic circuitry since the temperature of the components must not rise beyond acceptable levels that could damage them or affect their performance. The energy required for the cooling of data centers is generally quite substantial, with typical data centers requiring hundreds for megawatts for heat removal. Therefore, it is important to develop efficient thermal management systems to reduce power consumption and minimize the environmental impact. Extensive work has been done in recent years on the cooling of data centers. Though most studies have focused on steady state processes, time-dependent behavior is also of considerable interest because the thermal load is generally not constant but varies with time. Fluctuations in the load may arise or the load may peak at certain times. Efficient time-dependent distribution of load between various data centers, if several data centers are available to a given organization, or between different servers of a given data center may be used to improve the efficiency of the heat removal process and thus reduce the overall power consumption. Since environmental conditions vary from location to location, appropriate data centers may be selected, from those available, in order to use the most favorable location to reduce the overall cooling load. These aspects demand a study of both the steady and time-dependent behavior of data centers. The results obtained may be used to optimize the thermal management system and reduce the overall energy consumption. This paper reviews some of the relevant work done on data center cooling, with particular attention given to load distribution and to the data center location that determines the local environmental conditions. These results could form the basis for optimizing the cooling strategy for a system of data centers.

Keywords: *Data center, thermal management, optimization, transient effects, thermal load variation, environmental conditions*

1. INTRODUCTION

One of the major problems that has emerged in recent years is that of thermal management of data centers, which are employed extensively for data storage and retrieval. It is expected that the use of data centers and the need for accurate and dependable data storage will continue to grow in the future, placing significant challenges on thermal management. It necessary to study the cooling system for data centers since the temperature constraints on the electronic circuitry must not be violated for satisfactory performance, while minimizing the power consumption and the environmental impact. The cooling of typical data centers often requires considerable amounts of energy, very much like power plants (Patel, 2002; Joshi and Kumar, 2012; Khalaj and Halgamuge, 2017). There has been growing interest and research activity in this area, particularly for steady state operation of data centers (Chen, 2005; Rambo and Joshi, 2007; Choi, 2008; Zhang, 2008; Abdelmaksoud et al., 2010a; Patankar, 2010; US DOE, 2014; Gao et al., 2015). The focus in these studies has been on the flow configuration and operating conditions that would provide adequate cooling to meet the temperature constraints on the various components. For data centers, the constraints are generally specified on the average room temperature.

However, it is also important to study the effect of changes in thermal load and environmental conditions on the behavior of the system. Data centers are typically subjected to varying loads due to activity in storage and retrieval of data. Though some fluctuations around steady-state operation are expected, major changes can also occur over the day and over longer periods of time. Some of these variations are expected and may be predicted, whereas sudden unexpected surges may also arise. All these considerations make it imperative to determine the transient characteristics of data centers and design heat removal systems to meet the challenges posed by load and environmental variations.

Modeling and numerical simulation, along with selective experimentation, have been used to obtain the flow and temperature fields, as well as heat transfer rates in data centers (Iyengar, 2007; Amemiya et al., 2007; Abdelmaksoud et al., 2010b; Nada and Elfeky; 2016; Beghi et al., 2017). These results can be used to determine hot spots in the electronic hardware and areas that need heat transfer enhancement, so that the thermal management system may be designed to meet the stringent temperature constraints of electronic systems that typically require device or electronic chip temperatures to be lower than about 80 ^{o}C. Variations in flow configuration and operating conditions like flow rates and inlet temperatures, that are controlled by the use of chillers, may then be employed to obtain satisfactory thermal management. The design of the hardware consisting of electronic components, servers and racks may also be considered in some cases to improve the heat removal.

The air flow within the data center has a major effect on the temperature distribution in the equipment located in the rooms. In some data centers, cold air enters the data center from the ceiling through

* *Corresponding author. Email: jaluria@soe.rutgers.edu*

diffusers and exits the room via vents on the sides of the room. However, most current data centers use the hot aisle/cold aisle layout, as shown in Fig. 1. This arrangement is designed to supply cold air through a raised floor. Computer room air conditioning units (CRACs) are used to pump the cooling air into the plenum underneath the room. There are perforated tiles on the floor of the cold aisles that allow air to enter the space above the floor. The aisles without the cold air inflow are termed hot aisles. This arrangement is popular because of its flexibility and versatility. If the layout of the server racks is changed, the corresponding perforated tile locations can be changed easily so that the cold air can be delivered to where the hot rack is located (Kang et al., 2000; Karki, 2003; Patterson, 2008; Khalifa and Demetriou, 2011; Nada et al., 2016; Fulpagare, 2017). Different parameters and operating conditions have been considered in the literature for a wide variety of data centers (Barroso et al., 2019). The overall system is a fairly complex one and various numerical models have been developed to simulate the flow and heat transfer in data centers.

Fig. 1 View of a typical cold aisle in a data center.

System optimization involves choosing the hardware, particularly the configuration of the server racks in the data center and of the convective flow, as well as the operating conditions, such as the flow rate and inlet temperature, in order to minimize the energy consumed. Chillers are needed to decrease the inlet temperature below the ambient temperature. Substantial energy savings are obtained if chillers are not employed and ambient air is directly used for heat removal. This circumstance is often known as free cooling and has been of considerable interest due to the reduced costs and simplifications involved (Goiri et al., 2015). A consideration of the heat input due to the electronic system, possible hot spots and effectiveness of the circulating flow would allow the determination of when chillers are needed to ensure satisfactory performance of the data center (Tschudi, 2003; Litner, 2010). If chillers are not used, air handling systems that employ fans, ducts and vents are used, with the inflow of ambient air into the system.

The effect of the environment conditions on the heat removal is another interesting consideration. The environmental conditions, particularly the humidity and the ambient temperature, obviously affect the energy needed for the cooling of a data center. Also, minimization of the power consumption for cooling would also decrease the impact on the environment. Several data centers, spread out over the country or the world, are often available to many large organizations. It is then possible to distribute the load among data centers, depending on the environmental conditions, to minimize the power consumption (ASHRAE, 2008; Demetriou and Khalifa, 2011; Le et al., 2011; Zhang and Jaluria, 2017). Colder regions may be effectively employed in the summer and warmer ones in the winter without the extensive use of the power consuming chillers to cool the air entering the data center. The electronic load to a given data center could be kept low in order to cool the system with the use of a simple fan. For larger thermal loads, chillers will be needed. The load and location of the data center can thus be optimized in order to minimize the power consumption and also reduce the effect on the environment. Obviously, the transient behavior of data centers is needed to effectively vary the load distribution among different data centers and also to design the cooling system for a given data center for variations in load with time (Moore et al., 2005; Ghosh et al., 2011; Ghosh et al., 2011b; Erden, 2013; Fulpagare et al., 2016).

There are two classes of thermal management policies for data centers: those that manage temperature under normal operation and those that manage thermal emergencies. The objective for normal operation thermal management is to reduce the cooling cost. On the other side, a large increase in load that causes temperatures to rise quickly can be considered a thermal emergency. The main objective for managing thermal emergencies is to control temperatures while avoiding unnecessary performance degradation without excessive cooling capacity. Both these thermal management policies have been studied.

This paper reviews the results on the steady and transient operation of data centers under different environmental conditions. Changes in the thermal load are also considered. Different scenarios where the cooling system is started before the data center is subjected to a major load increase are considered and the corresponding results presented. The study could be used for optimization of multiple data centers, with respect to load and location, to achieve considerable savings in energy usage.

3. MATHEMATICAL AND NUMERICAL MODELING

As mentioned above, a wide range of data center configurations are in use, with different thermal management strategies and thermal loads that can be handled. Consider, for instance, the simple system shown in Fig. 2. It consists of rows of server racks, with 8 racks in each row. The overall dimensions of the room are 10m x 7m x 3m and each server rack is 1m x 1m x 2m tall. The cold air flow enters through the raised floor and the heated air exits from the top of the room through a plenum or grilles. The air is circulated in the data center by a set of CRACs positioned as shown. The CRAC units discharge cold air into the under- floor plenum and the air is delivered to the raised floor through perforated tiles, then the hot air returns to the CRAC units. Different levels of data center utilization, or thermal loads, are modeled by selectively turning the racks on or off. Because of symmetry, only half the region may be simulated to save computational time. This is obviously a fairly simple, though realistic, data center, which indicates the basic features that need to be considered. A somewhat different data center is considered later to present additional results.

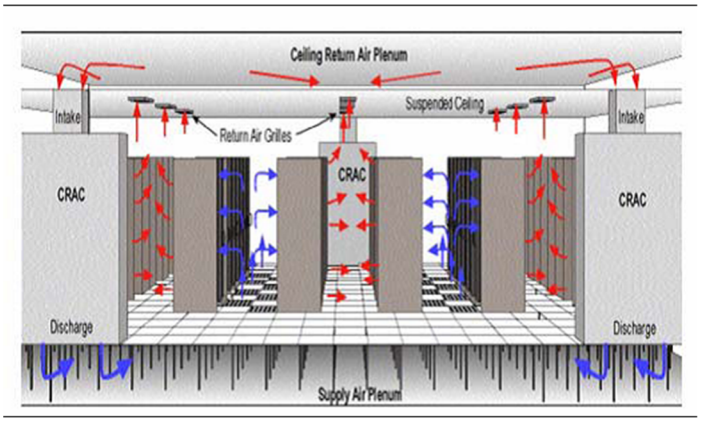


Fig. 2 A typical raised floor data center model.

3.1 Governing Equations

The flow in the data center is taken as turbulent, which instantaneously satisfies the Navier-Stokes equations as given below from Kundu and Cohen (2002):

$$\frac{\partial \tilde{u}_i}{\partial x_i} = 0 \tag{1}$$

$$\frac{\partial \tilde{u}_i}{\partial t} + \tilde{u}_j \frac{\partial \tilde{u}_i}{\partial x_i} = -\frac{\partial \tilde{p}}{\partial x_i} - g[1 - \beta(\tilde{T} - T_0)]\delta_{i3} + \nu \frac{\partial^2 \tilde{u}_i}{\partial x_j \partial x_j} \tag{2}$$

$$\frac{\partial \tilde{T}}{\partial t} + \tilde{u}_j \frac{\partial \tilde{T}}{\partial x_i} = \kappa \frac{\partial^2 \tilde{T}}{\partial x_j \partial x_j} \tag{3}$$

Here, u_i represents the velocity components, T the temperature, t the time, g the magnitude of the gravitational acceleration, x_i the coordinates, β the coefficient of volumetric thermal expansion and k the thermal conductivity. The physical dimensional quantities are indicated by tilde. It is challenging to predict the flow in detail since there are different scales to be resolved. Therefore, the averaged equations are used to find to mean velocity and temperature of a turbulent flow. Here, two equation models of turbulence are used in which the solution of two separate transport equations allows the turbulent velocity and length scales to be independently determined.

A two-equation κ-ω model, as well as a two-equation k-e model, are used in the simulations, yielding results that were fairly close. For the former, the equations for turbulent kinetic energy k and specific rate of dissipation w are employed. The equations for turbulent kinetic energy k and dissipation rate e are

$$\frac{\partial}{\partial t}(\rho\kappa) + \frac{\partial}{\partial x_i}(\rho\kappa) = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma_k}\right)\frac{\partial \kappa}{\partial x_j}\right] + G_k + G_b - \rho\varepsilon - Y_M \tag{6}$$

$$\frac{\partial}{\partial t}(\rho\varepsilon) + \frac{\partial}{\partial x_i}(\rho\varepsilon u_i) = \frac{\partial}{\partial x_j}[(\mu + \frac{\mu_t}{\sigma_\varepsilon})\frac{\partial \varepsilon}{\partial x_j}] + C_{1\varepsilon}\frac{\varepsilon}{k}(G_k + C_{3\varepsilon}G_b) - C_{2\varepsilon}\rho\frac{\varepsilon^2}{\kappa} \tag{7}$$

Here, G_k is the generation of turbulent kinetic energy due to the mean velocity gradient and G_b that due to buoyancy. Y_m represents the contribution of the fluctuating dilatation in compressible turbulence to the overall dissipation rate:

$$Y_M = 2\rho\varepsilon\frac{\kappa}{a^2} \tag{8}$$

In this study, air is treated as incompressible, so this term is zero. Also, $C_{1\varepsilon}, C_{2\varepsilon}, C_{3\varepsilon}$ are constants, σ_κ and μ_t are the turbulent Prandtl numbers for κ and ε, respectively. The corresponding values for the standard k-e model are 1.44, 1.92, 0.09, 1.0 and 1.3, as given in many references, such as Jaluria and Torrance (2003) and Minkowycz et al. (2006). Ansys Fluent 12.0 is applied with standard wall functions to solve the equations listed above. The SIMPLE algorithm is used to fully resolve the linear pressure-velocity coupling. The QUICK scheme is used to solve the convection-diffusion equations.

The grid dependence was studied for verification of the numerical model. The physical model was simulated with grids of different interval sizes. The results showed that the temperature and air flow distribution do not change beyond a particular grid size, which was chosen for subsequent simulations. The numerical models for the convective flow calculations were validated by comparisons with benchmark solutions on laminar and turbulent enclosure flows, particularly the 2D buoyancy-driven flow in a rectangular region and enclosure flows with inlet and outlet. Enclosures with isolated sources were also modeled and compared. The comparisons were found to be good, indicating the validity of the basic model.

The server rack is obviously a complicated computational domain since a wide range of configurations and server dimensions may be employed. Many different models have been proposed and applied to simulate typical server racks. If the configuration and placement of servers is known and fixed, appropriate models may be developed. However, given the diversity of configurations, a common approach has been to model the racks as porous media, with different porosities to represent the packaging employed in the system. Here, the server racks are simplified as porous media with a uniform distributed heat source, using typical data on configuration, component placement, energy dissipated, weight and volume. The perforated tiles are treated as one-dimensional porous jump boundary condition. The pressure drop across the perforated tiles is given by:

$$\Delta p = K(0.5\rho V^2) \tag{9}$$

where V is the velocity entering the perforated tile, and K is the flow resistance factor. It was calculated by the equations given by Idelchik (1986) :

$$K = \frac{1}{F}(1 + 0.5(1 - F)^{0.75} + 1.414(1 - F)^{0.375}) \tag{10}$$

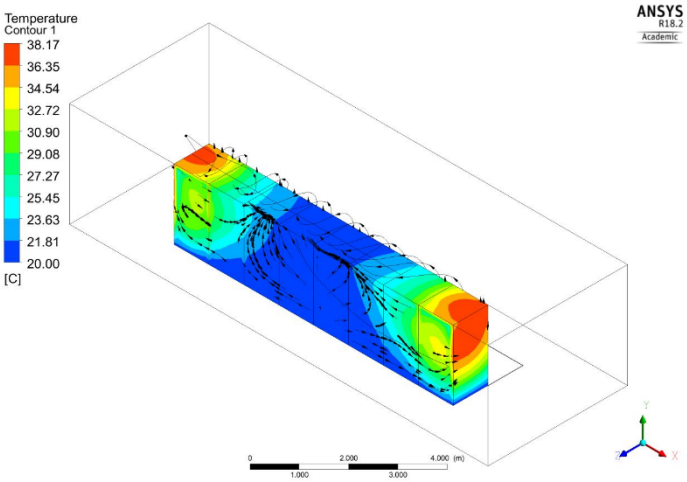
Here, F is the fractional open area of the perforated tile. Many other models are available in the literature for modeling the representative flow in porous media, both for the racks and for the perforated tiles. Higher accuracy is obtained by considering additional terms and mechanisms. However, this simple model adequately represents typical scenarios and allows a consideration of different aspects of the data center, focusing on steady-state operation, thermal load variations, and effect of environmental conditions.

4. RESULTS AND DISCUSSIONS

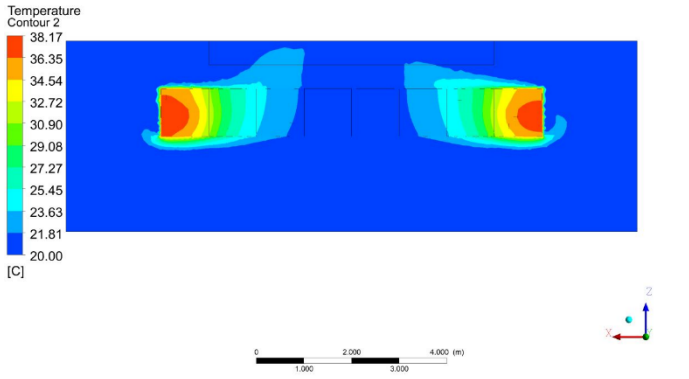
The effects of the operating conditions for different levels of data center utilization are studied. The small data center shown in Fig. 2 is first simulated here. Though a relatively simple model is used to approximate the server racks, a more detailed and accurate model can be employed if information on the system geometry, dimensions and components is available for a given data center. The data center utilization is not always 100 %. It varies depending on the need. Different cases of 25, 50, 75 and 100 % utilization are considered to determine the effect of the load on the flow and temperatures in the data center. For 25, 50 and 75 and 100 % utilization, 2, 4, 6 and 8 racks are in operation, respectively. Different scenarios of selective number of different racks in operation are considered. The thermal output from each rack is taken at a typical value of 10 kW for the results shown here. The porosity of the racks is taken at a typical value of 0.3; other values are considered later. The inlet velocity is given as 1 m/s or 1.5 m/s here. The inlet air temperature is taken as 20 ^oC without the chiller and as 12 ^oC with the chiller. However, the overall study considered a larger range of heat inputs and inlet temperatures and velocities. As mentioned earlier, only half the region is simulated due to symmetry (Zhang, 2012; Sunder, 2018). Therefore, the results are shown for a row of 8 racks, going from rack # 1 at the left corner to rack # 8 at the right corner. When scaling to larger data centers, a larger number of rows may be considered, along with different flow configurations. Symmetry may again be used in many cases to focus on one row of server racks.

Figure 3(a) shows the calculated temperature distribution and flow streamlines for 25 % utilization of the data center under steady conditions, with only racks 1 and 8 in operation and an inlet velocity of 1 m/s. Figure 3(b) shows the corresponding isotherms. Figures 4(a) and

4(b) show the results under the same conditions when only the middle racks, 4 and 5, are in operation. As expected, highest temperatures arise in the racks that are in operation. The maximum temperatures are seen close to the top since the air heats up as it moves from the inlet to the exit at the top. The streamlines indicate the flow in the racks, which are approximated as porous media. Air flows through the racks from the cold aisle to the hot aisle through the racks, while getting heated and removing dissipated energy from the servers. The racks heat up and lose energy by convection over the surface, particularly at the top surface due to the higher temperature there. A maximum temperature increase of only around 18 ^{o}C is observed because of the low thermal load. As the inlet velocity was increased, the temperatures decreased, as expected. It is also interesting to note that the second configuration, with the middle racks in operation, resulted in lower temperatures than when the outer racks were in operation. This is because hot air from the corner racks has to travel further and recirculate more within the system to reach the outlet, as compared to the middle racks. This can be seen from the isotherms that show the flow above the data center. Therefore, it is more efficient to turn on the middle racks first.

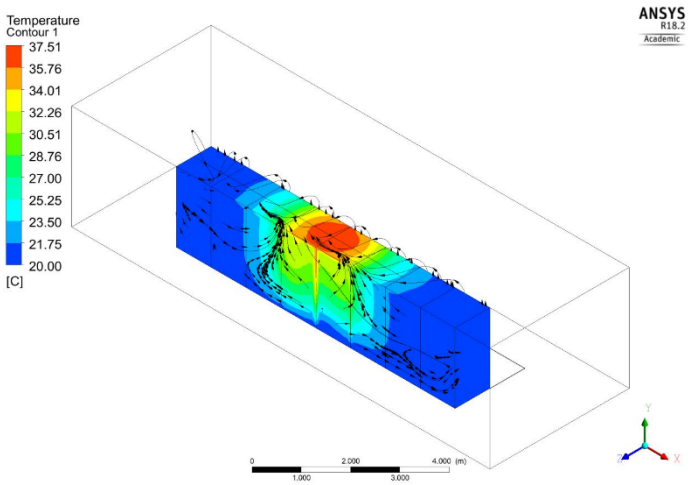


(a)

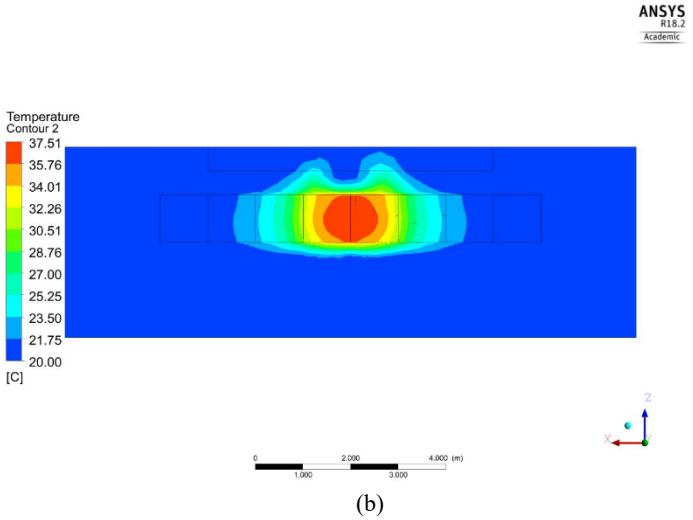


(b)

Fig. 3 (a) Temperature distribution and streamlines, and (b) isotherms, for 25% utilization with Racks 1 and 8 operating at 1m/s inlet velocity and 20 ^{o}C inlet temperature.



(a)



(b)

Fig. 4 (a) Temperature distribution and streamlines, and (b) isotherms, for 25% utilization with Racks 4 and 5 operating at 1m/s inlet velocity and 20 ^{o}C inlet temperature.

Figure 5 shows the results for 50 % utilization of the data center, with racks 1,2,7 and 8 in operation. As expected, the temperatures are higher than those for 25 % of the maximum load. The thermal effect is seen to spread out further from the corners and penetrate toward the middle because of the larger number of racks in operation. The flow above the data center is heated to higher temperature and is more vigorous due to buoyancy due to the higher heat input. From these calculations the local flow velocities and temperatures may be obtained and investigated for possible hot spots that may need adjustments in flow or rack configuration. It is seen that the hot spots that develop are concentrated when the middle racks are used and hence can be easily identified. Thus, they can be taken care of more easily by varying the server and flow configuration than when corner racks are used, which result in more dispersed hot spots.

Figure 6(a) shows the results for 50 % load, when the middle racks, 3,4,5 and 6, are in operation. As seen earlier, the temperatures are lower than when the corner racks are in operation. The thermal effect spreads out from the middle towards the corners and a strong buoyant flow is generated above the heated racks. The basic characteristics are the same as those seen in Figs. 3 and 4, except that the heat input is larger resulting in higher temperatures and stronger flow. The same configuration is considered for a decreased inlet temperature. With the chiller on, the inlet temperature is taken as 12 ^{o}C, keeping the inlet velocity at 1 m/s and

unchanged heat input. Figure 6(b) shows the results obtained. As expected, the temperatures decrease. The decrease in temperature follows an essentially linear variation as the inlet temperature. However, even though the decrease in inlet temperature cools the data center more effectively, it also consumes more energy and is more expensive due to the use of a chiller. Therefore, for given heat input, as indicated by the load, the middle racks may be turned on first and the inlet air temperature and velocity may be chosen to satisfy the constraint on temperatures. As mentioned earlier, the average room temperature in the data center is constrained so that the overall performance is satisfactory. Whether a chiller is needed or not will be dictated by the maximum temperatures attained in the room and if these values are acceptable. Also, if the ambient temperature is higher than the allowable room temperature, chillers will obviously be needed.

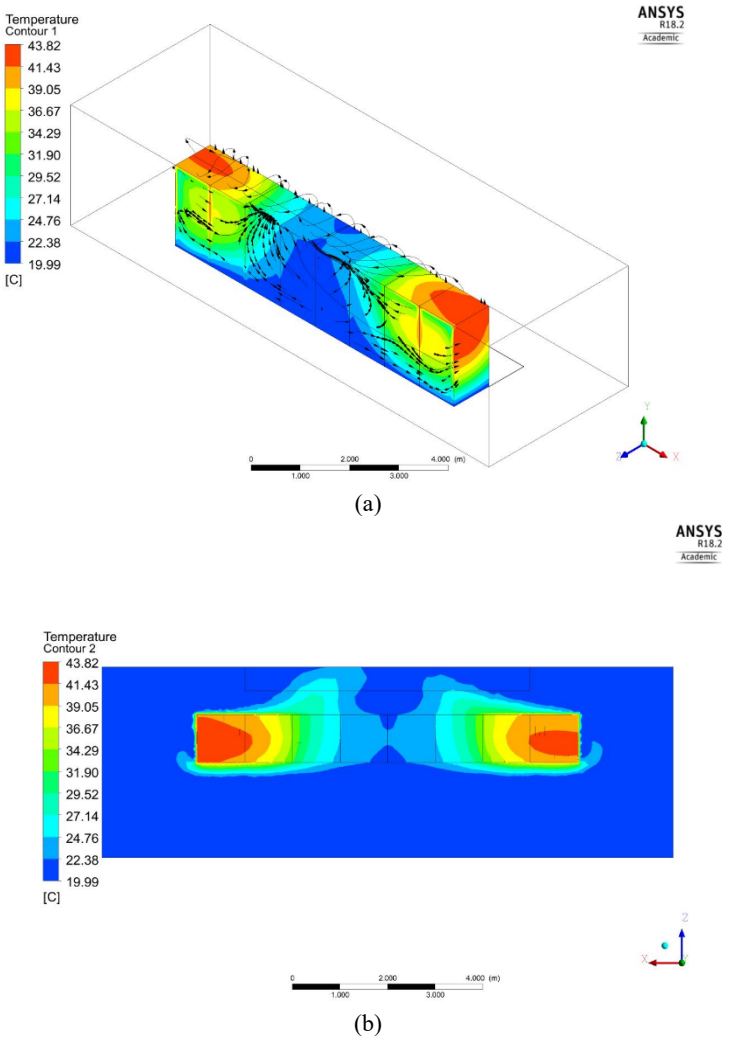


Fig. 5 (a) Temperature distribution and streamlines, and (b) isotherms, for 50% utilization with Racks 1, 2, 7 and 8 operating at 1m/s inlet velocity and 20 ºC inlet temperature.

Figure 7 shows the corresponding results at 12 OC air inlet for 75 % load. Six racks are turned on in this case. The temperatures increase, as expected, and the flow is more vigorous. The temperatures are lower when the middle 6 racks are turned on, rather than when the two middle ones are not in operation. Figure 8(a) shows the corresponding results when all the racks are turned on at 1 m/s air inlet velocity. The observed trends are similar to those discussed earlier. Figure 8(b) shows the results

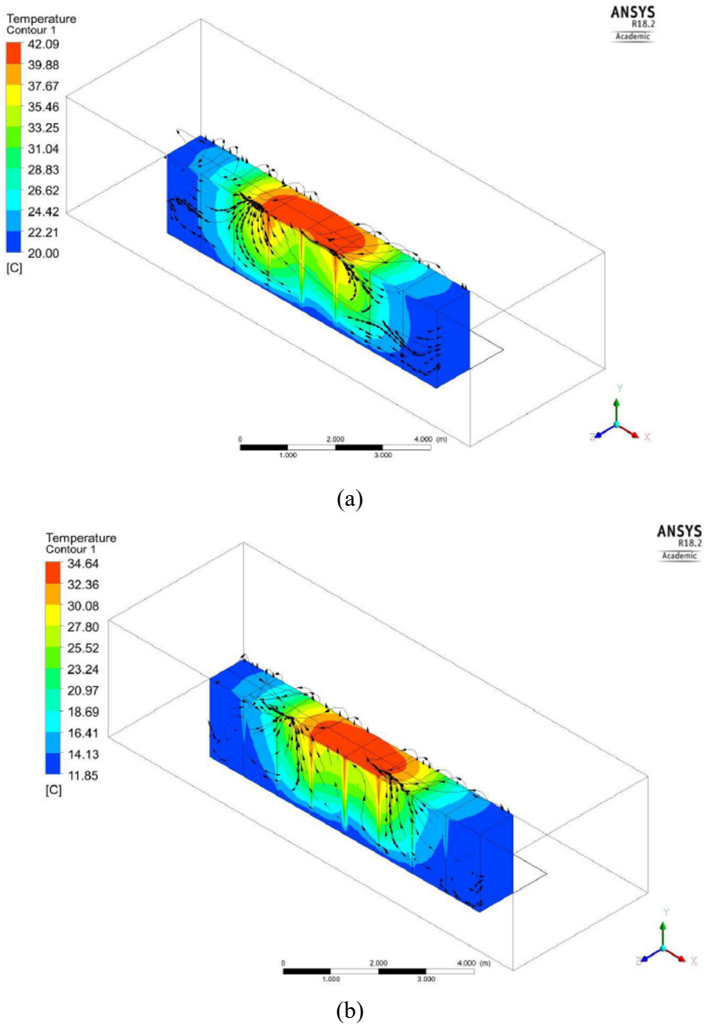


Fig. 6 Temperature distribution and streamlines for 50 % utilization with Racks 3-6 in operation, at inlet velocity of 1 m/s and inlet temperature of (a) 20 ºC and (b) 12 ºC.

when the inlet velocity is raised to 1.5 m/s. As expected, the temperatures decrease. The enhanced cooling, particularly in the middle of the row, is clearly seen. Therefore, depending on the thermal load, as determined by the data processing needed, different racks may be turned on for efficient cooling. The operating conditions, including the use of a chiller, are chosen on the basis of the temperature constraints and maximum temperatures that arise in the system.

The local temperatures, hot spots, air temperature distributions, and flow velocities can be obtained from the sample of results shown and discussed here. A few locations were chosen in the row of server racks to monitor the local temperatures, as shown in Fig. 9(a). Data exists for all the points in the chosen computational grid, but these chosen points reflect the general trends. For instance, the effect of increasing the air inlet velocity was seen in decreased temperatures and increased flow velocities. Figure 9(b) shows this behavior quantitatively at all the chosen sample locations for 100 % utilization. The temperature difference between all probes is not uniform since the heat input from each rack affects the adjacent rack and the air flow in and around the rack is not uniform. It must be mentioned that the results presented have focused on the rack temperature distribution. But a particularly important consideration is the room temperature. According to ASHRAE (2008) standards, the average room temperature must not be greater than 32 oC

in order to ensure that the data center performance is not adversely affected.

The racks, which are modeled as porous media, have a porosity, which depends on the packaging and dimensions of the servers. A range of porosities are considered to study the resulting effect on the data center heat removal. Figure 10 (a) shows the results for 50 % utilization at a porosity of 0.5 and Fig. 10(b) those at porosity of 0.7. It is seen that temperatures are lower with higher porosity This behavior has also been observed in practical systems. If the number of servers is reduced, even if the power output remains the same, reducing the servers increases the empty space within the rack. This allows more area for the air flow and hence allows more effective cooling. It can also be generalized that, for data centers which do not operate at full load, the racks can be designed to increase the open area by reducing the number of servers and hence increase the efficiency of the air conditioning system. Figure 10(c) shows these trends for 100 % utilization in terms of the temperatures at the chosen sample points. The expected behavior is indicated.

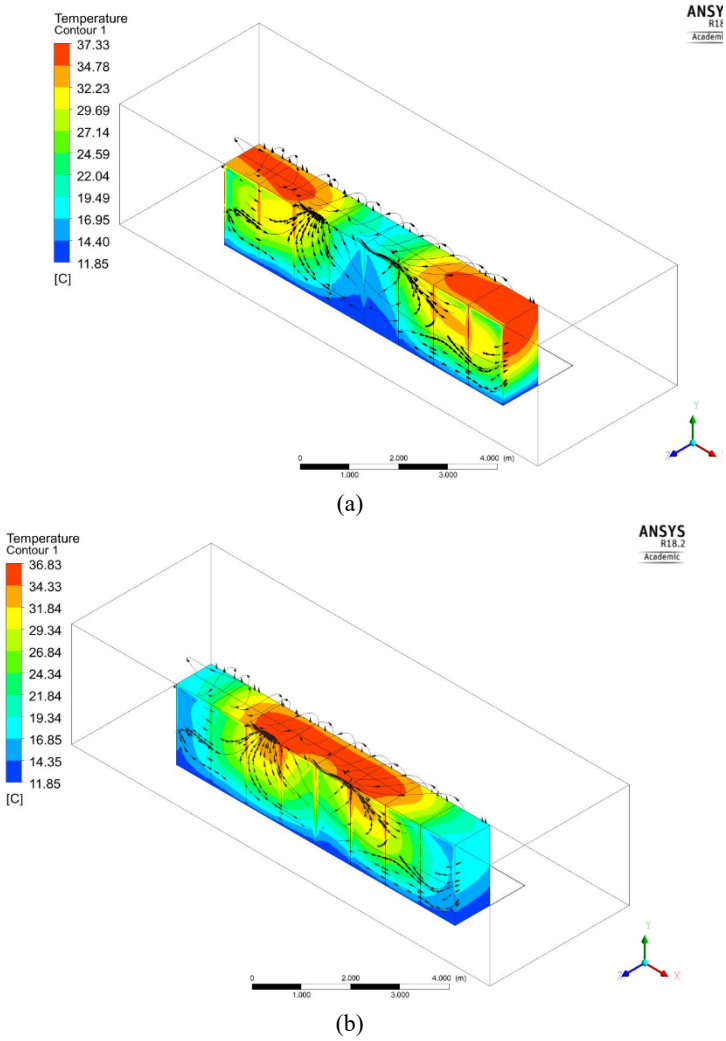


Fig. 7 Temperature distribution and streamlines for 75% utilization with (a) Racks 1, 2, 3, 6, 7 and 8 operating and (b) Racks 2-7 operating, at 1m/s inlet velocity and 12 ºC inlet temperature.

The preceding results indicate the basic features of thermal management of data centers under steady-state conditions. The temperature and flow distributions in the data center are obtained for a range of parameters and operating conditions, particularly thermal load. Similarly, results may be obtained for larger data centers, different configurations, and a range of operating conditions. All such results may then be used to optimize the cooling system and operate the data center with minimum energy input and reduced environmental impact.

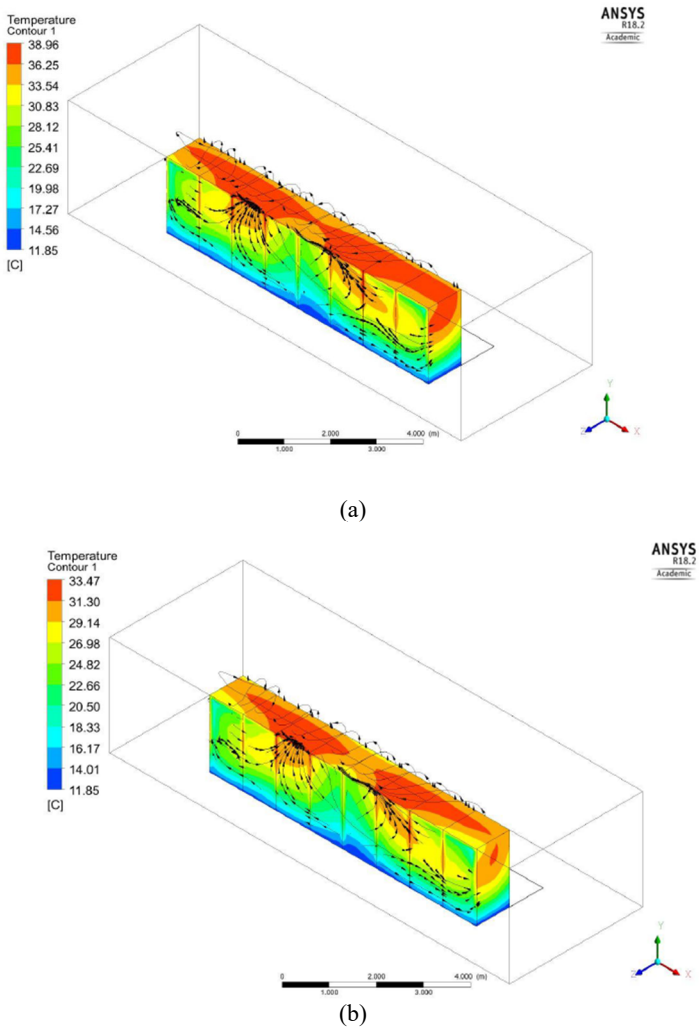
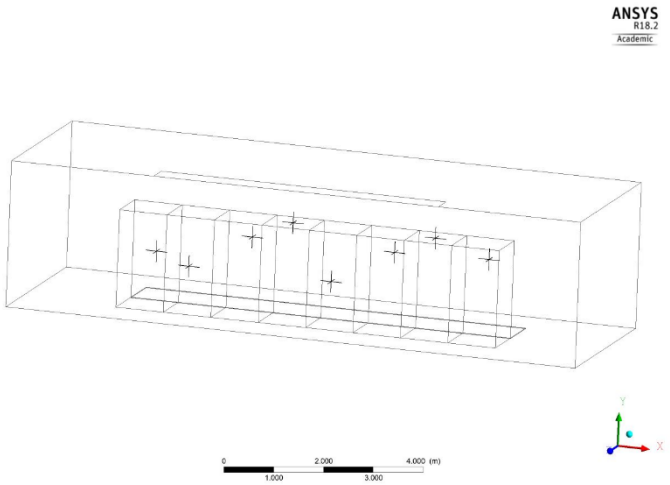
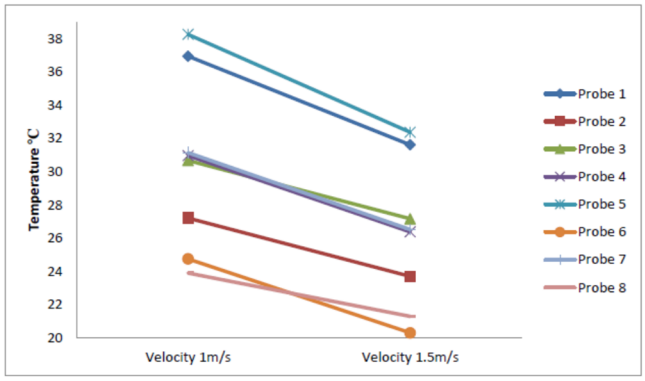


Fig. 8 Temperature distribution and streamlines for 100 % utilization at inlet velocity of (a) 1m/s and (b) 1.5 m/s and 12 ºC inlet temperature

Another data center is considered to further discuss the effects of thermal load and environmental conditions. The physical model is shown in Fig. 11. This data center is similar to the one considered earlier but has a break in the middle of the rows for better cooling characteristics. Also, the hot air exits from the sides where the air conditioning units are located. Symmetry is assumed by turning on the same racks in each row and the computational domain is taken as a quarter of the entire room. This data center consists of 6 CRAC units and 16 electronic racks (each 1m×1m×2m tall). The overall dimension of the data center is 7m× 8m×3m. The under-floor plenum height is 0.5m. Each rack is assumed to contain 24 servers, and the dimension of each server is 0.43m× 0.22m×0.046m (17in×8.5in×1.8in). Each rack has a power of 10 kW when fully utilized. Different combinations of racks are considered to be in operation for different loads. For a 25 % utilization, only one rack in each row of 4 racks is in operation. These are represented by Rack A_i, where i = 1, 2. 3 and 4. Therefore, the 4 corner racks are in operation. Similarly, other combinations are considered for different loads. The air flow rate is also varied. A larger number of rows of server racks may be considered to simulate larger data centers.



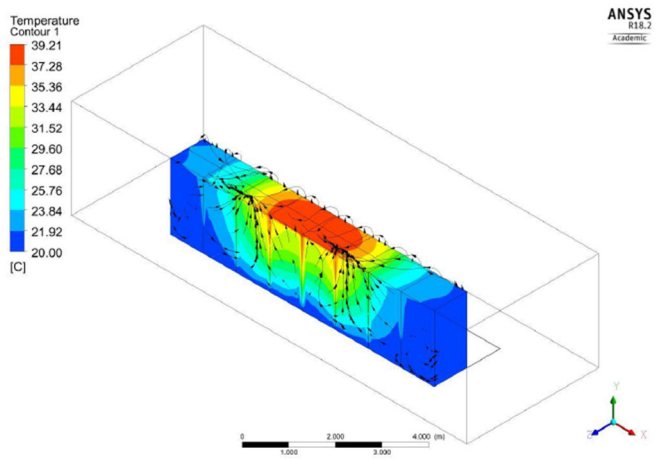
(a)



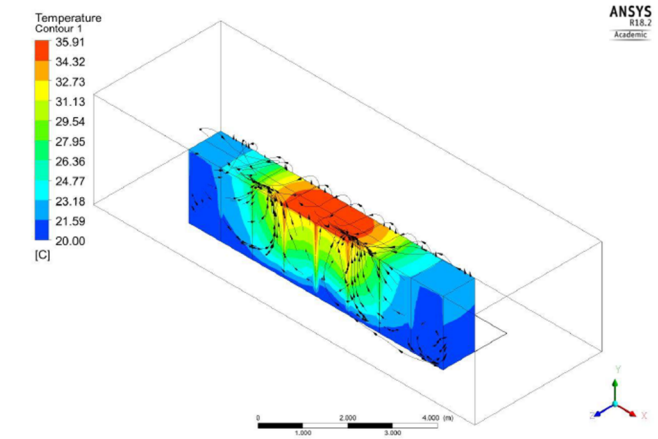
(b)

Fig. 9. (a) Probes for monitoring temperature changes (Nos. 2, 3, 7, 1, 8, 4, 5 and 6 from left); (b) Probe temperatures at two inlet velocities.

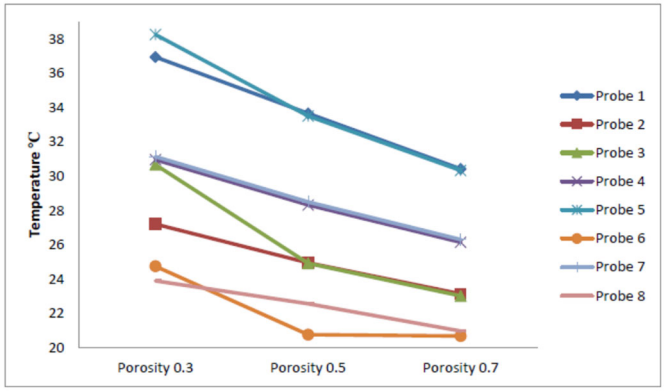
Some typical steady-state results are shown in Fig. 12. The inlet temperature is taken as 20 ^{o}C. The observed trends are similar to those seen and discussed earlier. The streamlines show strong flows near the corners and in the gap between the rows, thereby provide efficient heat removal in these regions. As expected, the temperatures are highest in the racks that are in operation. The highest temperatures tend to be near the top surface of the heated racks. Higher loads lead to higher temperatures. Racks Ai are found to be more efficiently cooled due to the proximity to the CRACs and thus shorter distance traveled by the air flow. The room temperature rises significantly as the thermal load is increased. Higher air flow rates lower the temperatures, as expected. Higher room temperatures, than those recommended by ASHRAE, have been used in recent years since the expected adverse effects have not been significant, depending on the design of the data center. The air flow rates can thus be adjusted if the room temperatures rise beyond the allowable limit for a given data center. It is also seen from these results that considerably higher loads can be satisfactorily managed by this data center without turning on the chillers. However, if the loads result in room temperature rise beyond acceptable levels, chillers may be employed to reduce the inlet air temperature to around 10 ^{o}C, or lower. Several different scenarios and operating conditions may be considered and the results obtained may be used for the design of the data center, as well as for choosing the operating conditions, particularly deciding if chillers are needed or not (Le et al., 2011; Zhang, 2012; Zhang and Jaluria, 2017; Barroso et al., 2019). Numerical results can also be used to pinpoint hot spots and make appropriate design changes to mitigate these.



(a)



(b)

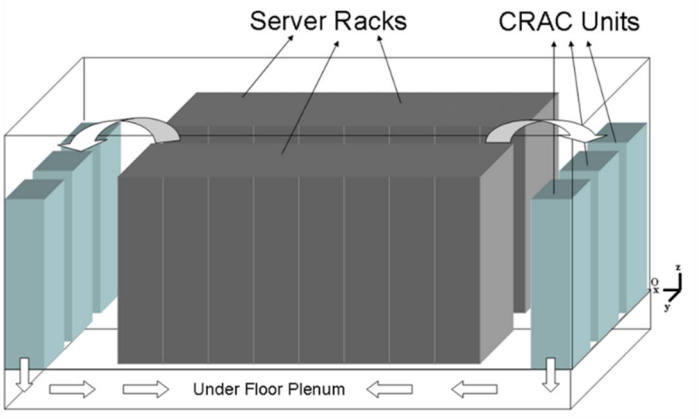


(c)

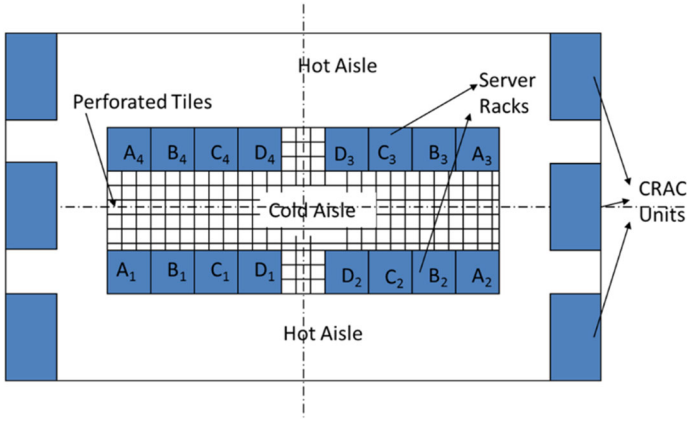
Fig. 10. Effect of porosity on the temperature and velocity distributions at 50 % utilization and inlet velocity of 1 m/s. (a) Porosity = 0.5; (b) Porosity = 0.7; (c) Variation of probe temperatures with porosity.

Another interesting consideration is the effect of the environment conditions on the thermal energy needed for cooling of a data center. In many cases, several data centers are available for a given organization, spread out over the country or the world. Then it is possible to vary the load, depending on the environmental conditions to minimize the power consumption (Abdelmaksoud et al, 2010a; Patankar, 2010). Thus, colder

regions may be effectively employed in the summer and warmer ones in the winter without the extensive use of the power consuming chillers to cool the air entering the data center. It is also possible to keep the electronic load low in order to cool the system with the use of a simple fan. At larger loads, a chiller will be needed. Therefore, the thermal load and location of the data center can be optimized in order to minimize the power consumption. For large data centers, heat removal poses many challenges for heat rejection and cooling towers, cooling ponds and other facilities are needed to get rid of the thermal energy. Therefore, an optimization of the thermal management process by controlling the load and by distribution of load among different data centers will also substantially reduce the effect on the environment. Figure 13 shows these effects quantitatively by considering two locations: Princeton and Seattle in USA. The former is considered in August and the latter in January, clearly resulting in significantly different ambient conditions. A higher load and higher ambient temperature demand higher power consumption, as expected. Information on the ambient temperature variation over the year at a given location may be used to vary the load for minimizing the energy needed for thermal management. Clearly, load distribution can be used effectively to lower the overall power consumption and the environmental impact. For further details on this aspect, the references given earlier may be consulted.

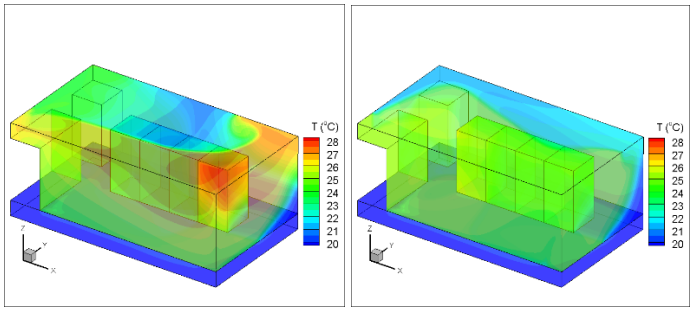


(a)

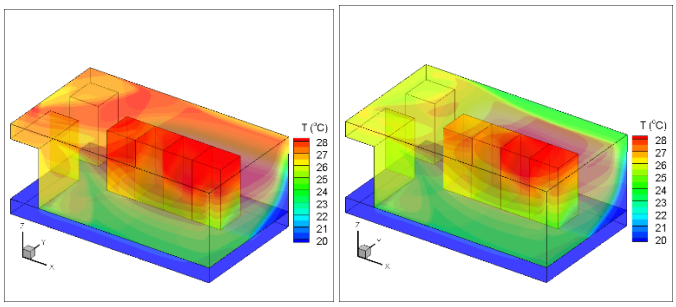


(b)

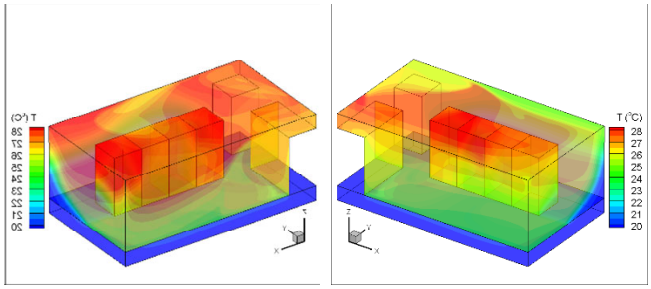
Fig. 11 Sketch of the data center considered, along with the CFD model used.



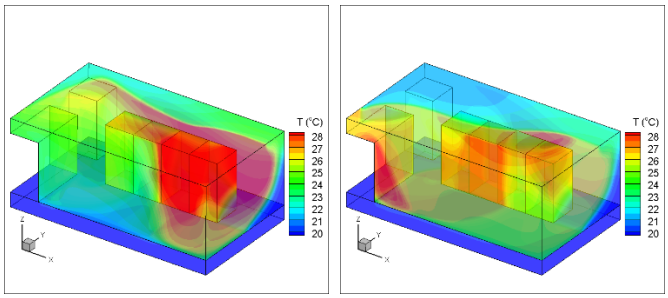
(a) D_i Racks in operation, 25% utilization , 2.195 m^3/s flow rate



(b) Racks A_i,C_i in operation, 50% utlization, 2.195 (b1) and 2.667 (b2) m^3/s flow rates



(c) 50 % utilization, 2.195 m^3/s flow rate. Racks A_i, D_i in operation (c1); Racks A_i,B_i in operation (c2)



(d) Racks A_i, C_i, D_i in operation, 75 % utilization, 4.814 m^3/s flow rate (d1). ; All Racks in operation, 100 % utilization, 5.663 m^3/s flow rate (d2)

Fig. 12. Temperature and flow distributions under various load conditions for different flow rates. The racks in operation and the flow rates are given in the figures.

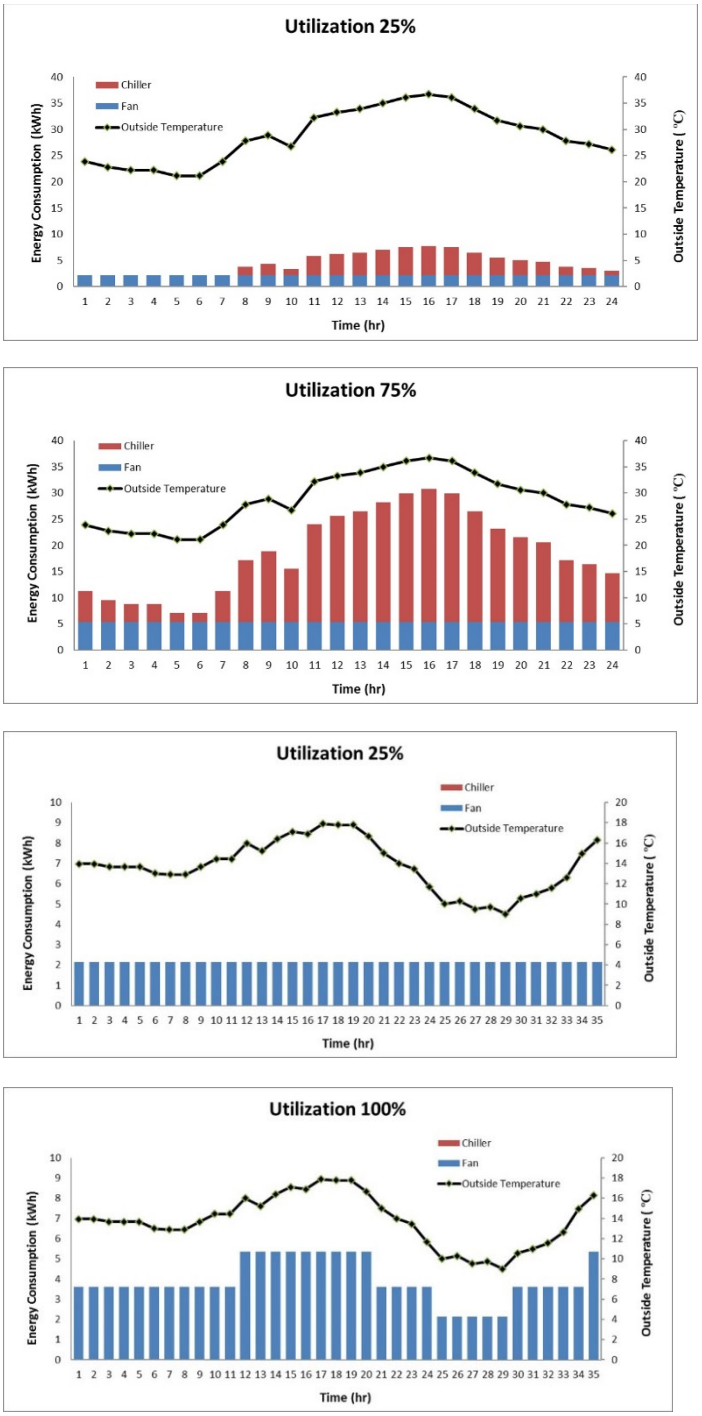


Fig. 13. Effect of environmental conditions and load on power consumption for cooling of data centers.Top two: Princeton, NJ, in August and Bottom two: Seattle, WA, in January. The red portion represents the power consumed by chillers.

4.1 Transient Effects

If there is a sudden increase in the thermal load from a steady situation, the cooling system must respond quickly and be able to handle the change. The air flow rate may have to be increased or a lower air supply temperature may be required to meet the increased demand. The system will experience a transient phase as it goes from one steady state to another. Similarly, the system will respond to fluctuations in load and to other load changes with time. The CRAC system has its own response time to be fully functional. This response time may be determined to a fair degree of approximation by using a lumped system analysis. For instance, if a Liebert Delux system/3 (A frame coil) is considered for the CRAC, the parameters like density, weight and volume of the cooling system can be employed to obtain the response time for different flow rates. Based on the numbers obtained from the lumped system analysis, the average response time for the CRAC system is obtained as around 20 minutes (Zhang, 2012). This response time can be considered along with the transient processes that govern the data center for computing the overall time-dependent temperatures.

Let us first consider the data center shown in Fig. 2. We had discussed the steady operation of this data center. Transient effects were also investigated for this data center. In typical cases, the load increases when there is a sudden surge in usage. This may result in a sudden rise in temperatures that may require adjustment of the operating conditions, as discussed earlier. Several scenarios of sudden increases in the thermal load were considered and one particular case is outlined here. The data center is assumed to be at 50 % load initially, with an inlet temperature of 20 ^{o}C and inlet velocity of 1.5 m/s. The load is then suddenly increased to 75 % with the chiller turned on to supply air at 12 ^{o}C and it stays at this level for 30 minutes, when it is again suddenly raised to 100 % with the other conditions remaining unchanged. The thermal response time of the chillers is taken as zero, i.e., immediate response. This is an idealization, used just to illustrate the transient effects in the flow and the racks. In practical systems, the response time of the CRAC will also be important, as discussed later.

The temperature distribution in the data center is monitored with time. There are two effects arising due to the increase in load and due to the inlet temperature decrease. The resulting effect of these two changes was found to show interesting response behavior of the different racks. Figure 14 shows the calculated temperatures at the 8 sample locations in the 8 server racks as functions of time. At time t = 30 minutes, the load changes from 75 % to 100 %. This is clearly reflected in the response of several probes. All the temperatures ultimately approach constant values to reflect the steady state conditions at large time. But some probes, particularly those near the edges, show a decrease in temperature with time. In regions where the airflow is relatively weak, the heat input results in a temperature increase. These points must be kept in check since they are potential hot spots. The use of the chiller results in temperature decrease in regions where the airflow is relatively strong. It is also seen that the overall response of the racks is quite slow since it takes a long time to reach steady state. These aspects indicate the possibility of fine tuning the response of the data center and the use of appropriate load distribution to obtain optimal cooling conditions. Efficient strategies are needed to keep the temperatures under the given constraints and improve the response of the system. This issue is discussed further later in the paper.

Let us now consider the second data center, shown in Fig. 11. Different cooling responses (scenarios) for sudden load changes occurring at t = 0 may be considered (Table 1). These are different strategies employed to meet the changes in the thermal load. Eight strategies are outlined here, with a brief description of each. Each strategy is a sequence of different actions taken, which include increasing the air flow, turning on the chiller or a combination of the two. As an example, let us consider the circumstance where the utilization is increased from 25% to 75% as a step change. The chiller is assumed to be off before the change. At a flow rate was 1.581 m^3/s, the temperature ramps up to 42 ^{o}C with nearly all the heat going into raising the temperature of the data center, as shown in Fig. 15. This is strategy 1. If the cooling level is increased at the same time, the temperature rise slowly decreases until a steady state temperature is again reached. This is strategy 2 in which the flow rate remains the same and the chiller is turned on at t = 0. The effect of the response time of the cooling system is included. The initial response of strategy 2 is the same as strategy 1; only that the steady state temperature is lower, as expected. Similarly, strategy 3, which aggressively increases the flow rate, stabilizes the

temperature at above 30 °C. Strategy 7, which ultimately uses a larger flow rate than strategy 3 but increases the flow rate more conservatively is not beneficial, allowing the temperature to rise over 32 °C before settling to a stable temperature slightly above 30 °C. Furthermore, note that for strategies 4, 5 and 8, which are different combinations of inlet flow rate and chiller input, the temperatures are either significantly above the allowable temperature of 30 °C (strategy 4), or slightly higher than the allowable temperature (strategy 5 and 8). The steady state temperature of all the three scenarios are under 30 °C. Apparently, it is not sufficient to turn on the chiller the very instant when the load changes. This will put the equipment under high risk and may cause a break down because of the high temperature.

To prevent the overheating caused by the response time required by CRAC systems, precooling may be employed. If the arriving thermal load exceeds the capacity of the current cooling system settings and may cause over heating problems, it could be on kept on hold or sent to other data centers. The control system may pre-cool the data center (by turning on the chiller early) in preparation for receiving load. The precooling time depends on the cooling scenario setting. As shown in Fig. 16, for strategy 4, even with 20 minutes precooling, the temperature will still rise to 32 °C. Strategy 5, which immediately increases the flow rate as the chiller is turned on, only needs to pre-cool for 5 minutes before the load change. For strategy 8, if the flow rate is increased gradually and meanwhile the system is pre-cooled for 10 minutes or longer, the temperature is controlled under 30 °C. It is unnecessary to pre-cool the data center for 20 minutes for strategy 8.

An important consideration in the design of the cooling system is the presence of uncertainties that arise in various parameters. Even if an acceptable design is obtained from deterministic models, the uncertainties can cause variations that can make the design unsatisfactory. Due to the existence of the uncertainties, the traditional deterministic formulation is no longer adequate to generate safe and acceptable designs because it may lead to a design with high risk of system failure. In order to achieve high reliability in the final design, it is necessary to develop reliability-based design, which results in failure rate lower than an accepted level, generally taken as 0.13%. The development of the reliability-based design optimization (RBDO) algorithm evaluates the probabilities of the system failures and provides a more conservative design which ensures that the failure probabilities are subject to some acceptable level. If any uncertainties are found in the experiments, the simulations, or the manufacturing process, the information on the uncertainties is fed back to the formulation of the RBDO problems, and new appropriate conditions can be generated. Several researchers have estimated the randomness of the operating parameters in different thermal processes. These parameters are then assumed to have a distribution of values, such as a normal distribution, rather than a deterministic fixed value. The design process then employs these distributions to obtain an acceptable design, with a range of variation in the parameters needed to meet the uncertainty level. This approach has been applied to a CVD reactor and to microchannel flows (Lin et al., 2010; Zhang et al., 2014). However, much more detailed effort is needed on this aspect to obtain realistic and useful designs of thermal systems.

5. CONCLUSIONS

The thermal management of data centers is considered in this paper. Both steady and time-dependent cases are considered, with the focus on variations in the thermal load and environmental conditions. Air cooling with cooled air entering the data center from under-floor plenum through perforated tiles is the cooling system considered here. Detailed results are reviewed for the steady operation of a couple of relatively simple data centers to indicate the main parameters such as the flow configuration and system design and the dominant operating conditions like inlet air flow rate and temperature. Different scenarios to achieve fractional utilization of data centers is investigated. The results indicate the need for chillers to lower the inlet air temperature when the ambient temperatures are high and when the load is high. At fractional load and relatively low ambient temperatures, chillers may not be needed, and fans may provide adequate cooling. Since free cooling, i.e., cooling with ambient air and without chillers is desirable for lower energy consumption and lower environmental impact, the conditions when the chillers may be avoided are discussed. The distribution of load among various data centers to achieve overall higher efficiency of heat removal is also discussed. Pre-cooling of the data center before the thermal load is sharply increased is discussed as a strategy to keep the temperatures within allowable limits. The use of these results for system design is discussed, along with possible uncertainties that may arise and practical issues that must be included.

ACKNOWLEDGEMENTS

The authors acknowledge the discussions with Professor Roberto Bianchini on Data Centers.

NOMENCLATURE

c_p	specific heat (J/kg·K)
g	gravitational acceleration (m/s^2)
k	thermal conductivity (W/m·K)
K	flow resistance factor
p	pressure (N/m^2)
Δp	pressure drop (N/m^2)
t	time (s)
T	temperature (K)
u_i	velocity components (m/s)
x_i	coordinate (m)

Greek Symbols

β	*coefficient of volumetric thermal expansion (1/K)*
κ	*turbulent kinetic energy*
ρ	density (kg/m^3)
ε	rate of dissipation of turbulence
ν	kinematic viscosity (m^2/s)

Superscripts

~	physical quantities

Subscripts

∞	ambient environment

REFERENCES

Abdelmaksoud, W.A., Khalifa, E., Dang, T.Q., Schmidt, R.R., and Iyengar, M., 2010a, "Improved CFD Modeling of a Small Data Center Test Cell," *12th Intersociety Conf. Thermal Thermomech. Phenom. Electronic Systems (ITherm)*.
https://doi.org/10.1109/ITHERM.2010.5501425

Abdelmaksoud, W.A., Khalifa, E., Dang, T.Q., Elhadidi, B., Schmidt, R.R., and Iyengar, M., 2010b, "Experimental and Computational Study of Perforated Floor Tile in Data Centers," *12th Intersociety Conf. Thermal Thermomech. Phenom. Electronic Systems (ITherm)*.
https://doi.org/10.1109/ITHERM.2010.5501413

Amemiya, Y., Hamann, H., Schappert, M., Van Kessel, T., Iyengar, M., O'Boyle, M., and Shen, J., 2007, "Comparison of Experimental Temperature Results with Numerical Modeling Predictions of a Real-World Compact Data Center Facility", *ASME 2007 InterPack Conf.*, **1**, 871-876.
https://doi.org/10.1115/IPACK2007-33899

ASHRAE, 2008, "Environmental Guidelines for Datacom Equipment," *American Society of Heating, Refrigeration, and Air-Conditioning Engineers*.

Beghi, A., Cecchinato, L., Mana, G.D., Lionello, M., Rampazzo, M. and Sisti, E., 2017, "Modelling and Control of a Free Cooling System for Data Center", *Energy Procedia*, **140**, 447-457.
https://doi.org/10.1016/j.egypro.2017.11.156

Barroso, L.A., Holzle, U. and Ranganathan, P., 2019, *The Datacenter as a Computer: Designing Warehouse-Scale Machines*, 3rd Ed., Morgan & Claypool Pub., San Rafael, CA.
https://doi.org/10.2200/S00874ED3V01Y201809CAC046

Chen, Y., 2005, "Managing Server Energy and Operational Costs in Hosting Centers," *2005 ACM SIGMETRICS Int. Conf. Meas. Modeling Computer Sys*., 303-314.
https://doi.org/10.1145/1071690.1064253

Choi, J., 2008, "A CFD-Based Tool for Studying Temperature in Rack-Mounted Servers," *IEEE Trans. Computers*, **57**, 1129-1142.
https://doi.org/10.1145/1071690.1064253

Demetriou, D.W. and Khalifa, H.E., 2011, "Energy Optimization of Air-Cooled Data Centers", *J. Thermal Sci. Eng. App.,* **2**, 041005.
https://doi.org/10.1115/1.4003427

Erden, H.S., 2013, "Experimental and Analytical Investigation of the Transient Thermal Response of Air Cooled Data Centers". Ph.D. Thesis, Syracuse Univ., Syracuse, NY, 2013.

Fulpagare, Y., Bhargav, A., and Joshi, Y., 2016, "Transient Characterization of Data Center Racks", ASME 2016 Int. Mech. Eng. Cong. Expo.
https://doi.org/10.1115/IMECE2016-66870

Fulpagare, Y., Joshi, Y. and Bhargav, A., 2017, "Rack Level Forecasting Model of Data Center". 16th IEEE Intersociety Conf. Thermal Thermomech. Phenom. Electronic Systems (ITHERM), 824-829.
https://doi.org/10.1109/ITHERM.2017.7992571

Gao, C., Yu, Z. and Wu, J., 2015, "Investigation of Airflow Pattern of a Typical Data Center by CFD Simulation", *Energy Procedia*, **78**.
https://doi.org/10.1016/j.egypro.2015.11.350

Ghosh, R., Kumar, P., Sundaralingam, V., and Joshi, Y., 2011, "Experimental Characterization of Transient Temperature Evolution in a Data Center Facility", 22nd Int. Symp. Transport Phenom.

Ghosh, R., Sundaralingam, V., Isaacs, S., and Kumar, P., 2011, "Transient Air Temperature Measurements in a Data Center", *10TH ISHMT-ASME Heat Mass Transfer Conf*., ISHMT_USA_013, 2011.

Goiri, I., Nguyen, T.D., and Bianchini, R., 2015," CoolAir: Temperature- and Variation-Aware Management for Free-Cooled Datacenter, *ASPLOS'15*, Istanbul, Turkey.
https://doi.org/10.1145/2694344.2694378

Idelchik, L.E., 1986, *Handbook of Hydraulic Resistance*, 2nd ed., Harper & Row, New York.

Iyengar, M., 2007, "Comparison between Numerical and Experimental Temperature Distributions in a Small Data Center Test Cell," *ASME 2007 InterPACK Conf.*, **1**, 819-826.
https://doi.org/10.1115/IPACK2007-33508

Jaluria, Y. and Torrance, K.E., 2003, *Computational Heat Transfer*, 2nd Ed., Taylor and Francis, New York, NY.

Joshi, Y. and Kumar, P., Eds., 2012, *Energy Efficient Thermal Management of Data Centers*, Springer, New York, NY.
https://doi.org/10.1115/IPACK2007-33508

Kang, S., Schmidt, R.R., Kelkar, K.M., Radmehr, A. and Patankar, S.V., 2000, "A Methodology for the Design of Perforated Tiles in Raised Floor Data Centers Using Computational Flow Analysis," *Proc. Inter Society Conf. Thermal Phenom. (ITherm)*.
https://doi.org/10.1115/IPACK2007-33508

Karki, K.C., 2003, "Use of Computational Fluid Dynamics for Calculating Flow Rates Through Perforated Tiles in Raised-Floor Data Centers," *Int. J. Heat., Vent., Air-Conditioning, and Refrig. Res.,* **9**, 153-166.
https://doi.org/10.1080/10789669.2003.10391062

Khalaj, A.H. and Halgamuge, S.K., 2017, "A Review on Efficient Thermal Management of Air- and Liquid-Cooled Data Centers: From Chip to the Cooling System", *Applied Energy*, **205**, 1165-1188.
https://doi.org/10.1016/j.apenergy.2017.08.037

Khalifa, H.E. and Demetriou, D.W., 2011, "Energy Optimization of Air-Cooled Data Centers", ASME. *J. Thermal Sci. Eng. Appl.*, **2**, 041005-041005-13.
https://doi.org/10.1115/1.4003427

Kundu, P.K. and Cohen, I.M., 2002, *Fluid Mechanics*. 2nd ed. Elsevier Academic, NY.

Le, K., Zhang, J., Meng, J., Bianchini, R., Jaluria, Y., and Nguyen, T.D., 2011, "Reducing Electricity Cost Through Virtual Machine Placement in High Performance Computing Clouds", *SC 11 Int. Conf. High Perf. Comput., Networking, Storage & Anal.,* Seattle, WA.
https://doi.org/10.1145/2063384.2063413

Lin, P.T., Gea, H.C. and Jaluria, Y., 2010, "Systematic Strategy for Modeling and Optimization of Thermal Systems with Design Uncertainties," *Frontiers in Heat Mass Transfer*, **1**, 013003-1-20.
https://doi.org/10.5098/hmt.v1.1.3003

Lintner, W., 2010, "Best Practices Guide for Energy-Efficient Data Center Design," *National Renewable Energy Laboratory (NREL),* U.S. Department of Energy.

Minkowycz, W.J., Sparrow, E.M., and Murthy, J., Eds., 2006, *Handbook of Numerical Heat Transfer*, Wiley, New York,

Moore, J., Chase, J., Ranganathan, P., and Sharma, R., 2005, "Making Scheduling Cool: Temperature-Aware Workload Placement in Data Centers", *Proc. USENIX Annual Tech. Conf.,* 61-75.

Nada, S.A. and Elfeky, K.E., 2016, "Experimental Investigations of Thermal Managements Solutions in Data Centers Buildings for Different Arrangements of Cold Aisles Containments", *J. Build. Eng.,* **5**, 41-49.
https://doi.org/10.1016/j.jobe.2015.11.001

Nada, S.A., Said, M.A. and Rady, M.A., 2016, "CFD Investigations of Data Centers Thermal Performance for Different Configurations of CRACS Units and Aisles Separation", *Alexandria Eng. J.,* **55**.
https://doi.org/10.1016/j.aej.2016.02.025

Patankar, S.V., 2010, "Airflow and Cooling in a Data Center", *ASME J. Heat Transfer*, **132**, 073001-1-073001-17.
ttps://doi.org/10.1115/1.4000703

Patel, C.D., 2002, "Thermal Considerations in Cooling Large Scale High Compute Density Data Centers," *8th Intersociety Conf. Thermal Thermomech. Phenom. Electronic System*, 767-776.

Patterson, M.K., 2008, “The Effect of Data Center Temperature on Energy Efficiency,” *11th Intersociety Conf. Thermal Thermomech. Phenom. Electronic Systems*, 1167-1174. https://doi.org/10.1109/ITHERM.2008.4544393

Rambo, J. and Joshi, Y., 2007, “Modeling of Data Center Airflow and Heat Transfer: State of the Art and Future Trends,” *Distributed Parallel Databases*, **21**, 193-225. http://dx.doi.org/10.1007/s10619-006-7007-3.

Sunder, A., 2018, “A Computational Study on the Steady and Transient Behavior of Data Centers,” M.S. Thesis, Rutgers University, Piscataway, NJ.

Tschudi, W., 2003, “Data Centers and Energy Use: Let’s Look at the Data,” *ACEEE 2003*, Paper#162.

U.S Department of Energy, 2014, “Data Center Energy Consumption Trends,” Retrieved 2014-6-20, http://energy.gov/eere/femp/data-center-energy-consumption-trends.

Zhang, J. and Jaluria, Y., 2017, "Steady and Transient Behavior of Data Centers with Variations in Thermal Load and Environmental Conditions," *Int. J. Heat Mass Transfer*, **108**, 374-385. https://doi.org/10.1016/j.ijheatmasstransfer.2016.12.028

Zhang, X., 2008, “Effect of Rack Modeling Detail of the Numerical Results of a Data Center Test Cell,” *11th Intersociety Conf. Thermal Thermomech. Phenom. Electronic Systems,* 1183-1190. https://doi.org/10.1109/ITHERM.2008.4544395

Zhang, J., 2012, “Cooling of Electronic System: From Electronic Chips to Data Centers,” Ph.D. Thesis, Rutgers University, Piscataway, NJ.

Zhang, J., Lin, P.T., and Jaluria, Y., 2014, "Design and Optimization of Multiple Microchannel Heat Transfer Systems," *J. Thermal Sci. Eng. Appl.,* **6**, 011004-1 to 10. https://doi.org/10.1115/1.4024706