Commercial Packaging Solutions for a Research Oriented Graphics Supercomputer

Kurtis Keller & John Poulton
University of North Carolina, Computer Science Department
Chapel Hill, NC 27599-3175, USA

Abstract

PixelFlow is a scaleable graphics supercomputer that is expected, on its completion mid-1995, to deliver image generation performance 10-20 times that of present-day 3D graphics systems. Achieving this level of performance will require simultaneous solutions to several difficult problems in thermal management, power supply, and high-speed interconnections. This paper describes our solutions to these problems, specifically: a thermal management system that handles unusually high heat density with conventional forced-air cooling; a UL-approvable distributed power supply that provides the required very high operating currents to the system's daughter boards; and a novel packaging approach featuring daughter cards plugged into a mid-plane, which helps both in shortened signal connection distances and in power distribution.

1. Introduction

Several generations of experimental high-performance computer graphics systems have been built in the Microelectronics Systems Laboratory in the Computer Science Department at the University of North Carolina. Our latest completed system, Pixel-Planes 5, became operational in 1990. It has demonstrated speeds of over 2.3 million polygons per second and has been described by *Computer Graphics World* as the "world's fastest graphics computer."[1] Since the machine was a research prototype, its implementation took advantage of design practice that would not be acceptable in a commercial machine: it must be housed in a machine room with continuous 58°F (14° C) cooling air, it has no EMI shielding; its multiple 5kW card cages are cooled by noisy trays of unshielded 6" (15 cm) tube-axial fans; and it made little use of DFM techniques.

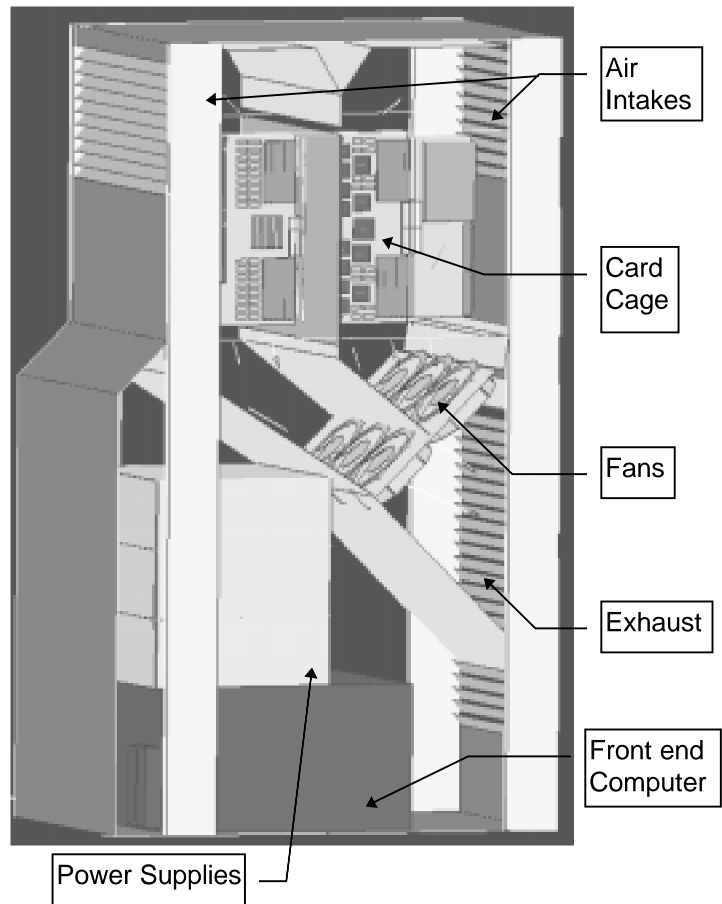

Figure 1

Our new generation machine, called PixelFlow, is expected to deliver scaleable performance 10 - 20x that of Pixel-Planes 5. To achieve these very high image generation rates required a distributed (point of load) power supply design and a cooling system able to remove over 590 watts per board set. PixelFlow will also be commercialized, so it must operate reliably at standard office environment ambient temperatures, generate relatively little noise, meet FCC and UL requirements, and require no fluid or external cooling apparatus.

[1] Cramblitt

2. System Packaging

PixelFlow's main system interconnection consists of some 544 point-to-point signals that daisy chain from board to board. This interconnection uses a very aggressive approach to obtain over 100Gbits/sec of bandwidth from board to board by using simultaneous bi-directional signaling, very low-voltage logic swings, and 200MHz signaling rates. This signaling system requires carefully controlled impedance, minimum possible cross-talk, and very short interconnection distances from card to card. These very short interconnect distances causes severe packaging problems that greatly restrict power distribution, board layout, cooling and board and component yield.

To solve this difficult problem, we use an unusual variation on the standard backplane/motherboard arrangement; the backplane is mid-plane, located along the center-line of the system cabinet (Figure 1 overall and Figure 2 for board detail). On the left hand side is half of the PXFL board set called the "Controller." It contains the two 40 watt floating point processors, their memory, three power supplies and miscellaneous house keeping circuitry. This 300 watt board is connected by a 385 pin connector to the backplane through which 95 of the data pins are passed directly to the "Render" card on the other side of the backplane. The remaining pins are connected to the backplane for power, grounds, and inter-board communications.

The "Render" board is located in-line with the Controller board but on the right side of the backplane. Unlike the simpler 8 layer Controller card, the Render card is a 14 layer, ultra high density board with 80 main active components, surface mounted chips mounted side by side on both sides of the board, a 350 and a 1225 pin backplane; two full size power supplies, five buses (including 512 and 128 bit wide 200Mhz paths); all on a 350 x 200mm PC board. Of the chips on the Render board, 76 produce no less than 4 watts each, while 12 are over 7 watts each. All of the high power chips have appropriately sized heatsinks

Video and I/O daughter boards plug into the auxiliary 350 pin connector on the right end of the Render board. Each of the video daughter cards has its own power supply for clean data conversion, and the card format provides enough space for a 3 1/2" hard drive. We anticipate designing a variety of video I/O daughter cards for various special purposes. These can be easily removed and replaced by the user without the need for tools. The Controller and Render cards will require a screwdriver for removal, partly to restrict their replacement to service personnel.

Interconnections between the boards and backplane were severely restricted by many constraints, including: the large (1225) pin count backplane per Render board; the requirement to bus 48VDC safely to the cards at under 8 amps each; impedance matching, cross-talk reduction, and signal length minimization needed to maintain signal integrity in the backplane interconnect; high pin density; requirement for press-fit assembly from both sides of the backplane; and low insertion force necessary to make the large pin count feasible. AMP Corporation's Z-PACK 2 mm HM interconnection system (under the IEC 1076-4-00x standard) was one of the few high density systems that could meet our unique requirements. This connector series also reduced the cost of the backplane and connectors themselves by allowing the tails of the backplane connectors for the Render board to extend through the backplane and be shared by the Controller board on the other side of the backplane.

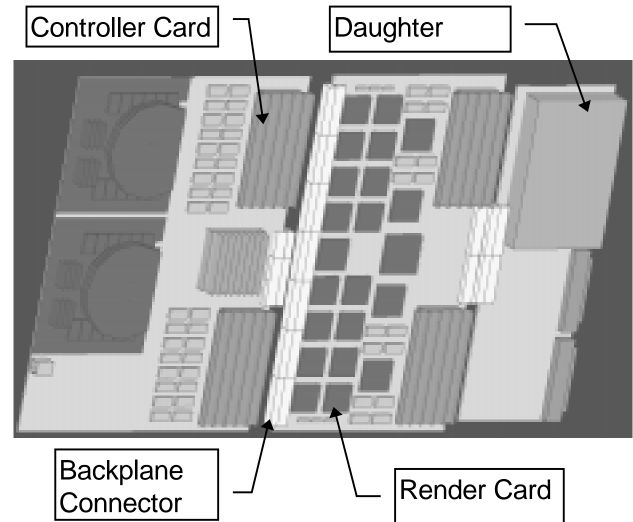

Figure 2

The overall system enclosure is shown in Figure 1. It includes the main PXFL card cage, which supports the mid-plane and daughter cards; below that is a standard 19" equipment rack on the left that houses the two 48VDC front-end power supplies, all monitoring equipment, the host computer, and other ancillary video, audio, and data storage equipment, typically tailored for an individual site installation. The six-fan assembly is located at a 45 degree angle

on the lower backside of the unit, pulling air through the top and exhausting out the lower back side.

PixelFlow can be scaled to sizes that exceed a single card cage frame. The high-speed point-to-point interconnect must be able to span multiple cabinets. These connections are made by means of a very short flexible circuit board with connectors that mate to the back side of the connector pins of the end daughter boards (this interconnecting board carries signals only; power is local to a card cage). The flex board is used to compensate for minor mechanical misalignments between individual racks.

3. Power Distribution

Each PixelFlow equipment rack contains 16 printed circuit board sets each containing two 40-Watt or 80-Watt RISC processors, a graphics raster processor with 44 power intensive custom chips, 64 Synchronous Dynamic RAMs (SDRAMs), 5 on-board power supplies, a daughter board and a large number of other components. All together, about 590 watts has to be dissipated on each of the board sets. Each board set requires 135 amps of 3.3V in addition to +5V at 14 amps, +12V and sometimes +4.4V. At these power levels, board-mounted point-of-load 48 volt, distributed power supplies appear to be the best solution. Ultra clean power for the video daughter boards will be supplied on the daughtercards themselves.

A major obstacle in powering each of these board sets was in complying with Under-writers Laboratories safety specification (UL 1950), which limits current levels for low voltage systems to 8 amps. We exceeded this limit by more than 50% in our first cut at the design, which used a single printed circuit board. Dividing the PXFL unit into two distinct boards that share data and power pins directly through the backplane allowed for power requirements per board to drop to a safe level, and allowed for much easier upgrading of the system. This approach also reduced the overall cost of the boards and assembly by having the ultra high speed and high density components on one small board and spreading out the medium speed (120 MHz), medium density section on a second, single sided, low layer count (cheaper) board.

Since the backplane is actually a mid-plane (Figure 3), the 48VDC from the front end power supplies to the backplane is interconnected by plug-in connectors at the bottom of the backplane board. The backplane is designed to meet UL requirements for spacing between traces of primary and secondary circuits. In addition, each 48V connection to a card is fused to the backplane for additional safety.

A system monitor oversees the system power and cooling functions, including startup and shutdown sequences, and continuous monitoring of power, temperature, and airflow. To control the PXFL system, a controller oversees all aspects of the system from startup and shut down to temperature and power monitoring while the system is operating. The monitor continually measures the 3-phase power input voltage and "voltage-good" signals from the point-of-load supplies on each board in the system.

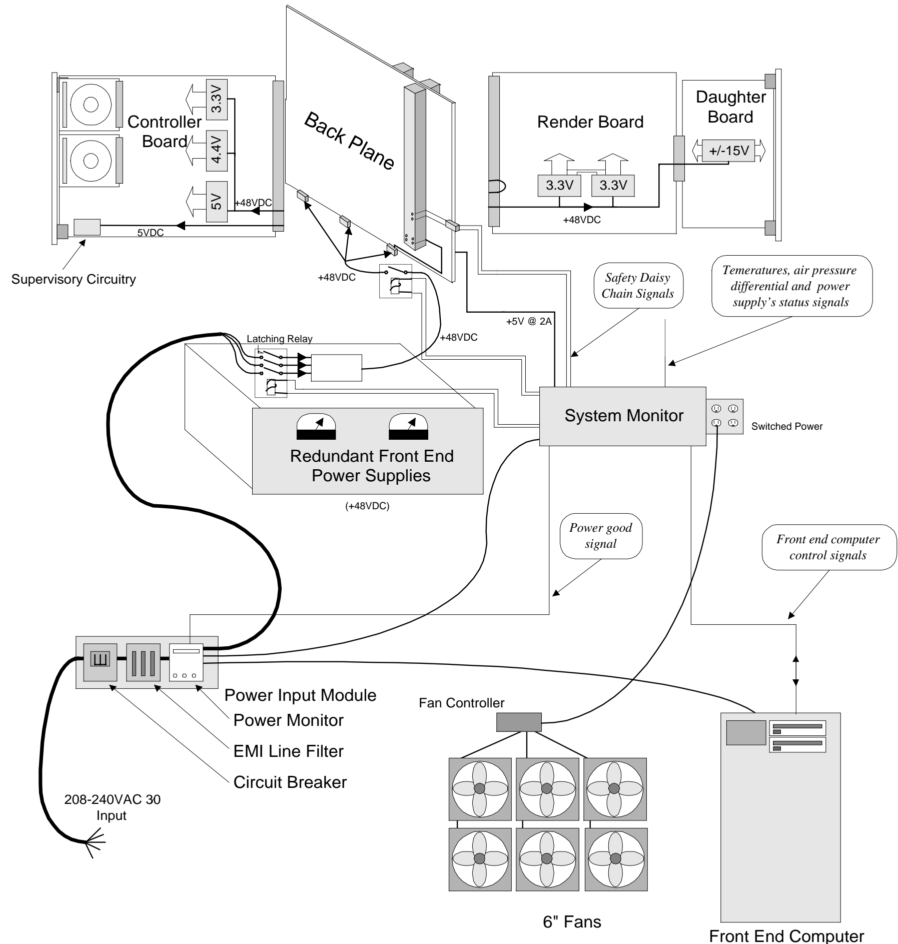

Figure 3

In the event of a failure, the monitor shuts down the entire system and presents error information for use by repair personnel. The system monitor also measures the temperature of each point-of-load supply and the exhaust air temperature. This information, along with temperatures measured in other parts of the system, allow the controller to adjust fan speed for constant operating temperature, thus ensuring higher system reliability, and lower noise.

4. Cooling

For simplicity, ease of maintenance, and low cost, forced air was the best choice for our cooling system. This design was difficult to implement, however, because of the high power density in the system. Though the system boards are mounted on fairly generous 35mm (1.4”) centers, ensuring adequate airflow was complicated by the large number of high-dissipation chips with heat sinks, mounted on both sides of the board, as well as point-of-load supplies, also equipped with large heat sinks. In addition to its on-board power supplies, the Controller board also contains two high-dissipation processor modules with integral heat sinks, 25 mm tall and 75 mm in diameter.

To ensure adequate cooling required considerable analysis of the air flow across the boards and through the system enclosure. Figure 4 shows a cutaway of air flow through the system, with air pulled in from the top, through the cards, through the fan array, and out the bottom rear of the unit (units in Figure 4 are in/sec). Figure 5 shows a detail of the intake to the card cage, revealing turbulence near the center of the cage; based on this simulation, baffles were added to reduce turbulence. Many iterations of the design were made, based on these useful fluid finite-element analysis simulations, gradually refining the design of baffles and fan placement. Once a reasonable approximation for bulk flow and temperature rise was achieved, we performed similar simulations of airflow across the boards and chips to determine case and junction temperatures.

The cooling system design was complicated by other factors. First, the system enclosure is about 1.5 meters tall, so it is possible for users to place items such as coffee cups on its top; the enclosure must therefore be designed to prevent spills from entering the cabinet. Second, since the system may be located in an office environment, it must be designed to reduce noise as far as possible and to avoid blowing exhaust air into the faces of users and service personnel.

The solution we have described, with air drawn in at the top and exhausted at the bottom, offers the advantages of a closed top (to prevent spills from entering the cabinet) and exhaust air directed toward the floor. Placing the fans on the downstream side keeps airflow through the system laminar, thereby reducing back pressure and unwanted eddies around components. Some turbulence is required around active components, to promote efficient heat transfer, and heatsinks are provided for this purpose.

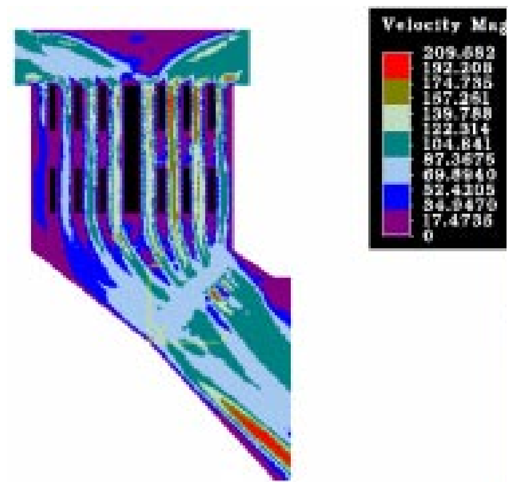

Figure 4

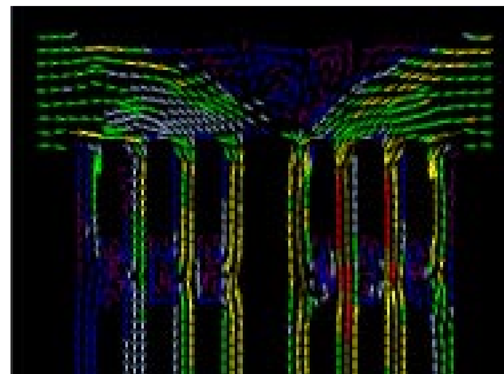

Figure 5

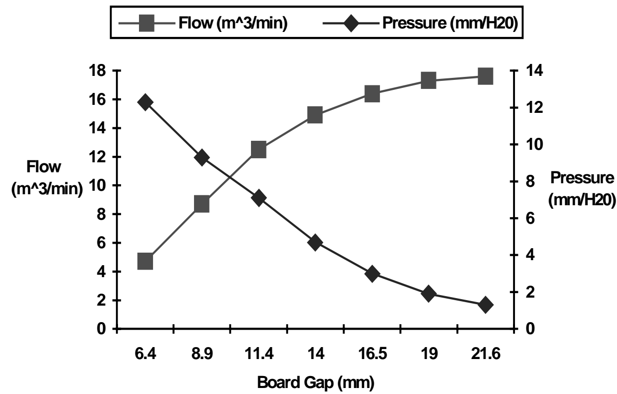

Figure 6

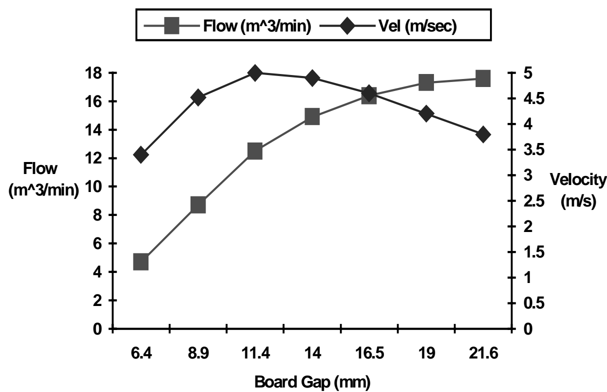

Figure 7

Petukhov Equation:

$$Nu=\frac{Re*Pr}{X} (f/8)*(\mu_b/\mu_w)^n$$

System airflow was analyzed using a 3D fluid flow package as follows: First, a 3D simulation of airflow between the circuit boards was performed, in which the backpressure from fan performance curves was matched with airflow across the boards. Next, this data was brought into a 2D F.E.A. fluid flow package, and the backpressure and air velocity were simulated for the card cage. Finally, the change in overall backpressure was brought back into the 3D card cage analysis for recomputation. This design loop continued until equilibrium was achieved. A full 3D analysis of the cards and the enclosure was attempted, but abandoned because of computing time and memory requirements (the 2D analysis alone required over 50 Mbytes of storage).

An accurate prediction of chip die temperatures required further detailed analysis of the backpressure caused by roughness of the chips and their heatsinks. Simulations using Petukhov's heat transfer equation for rough pipes for higher air flow rates would use the ratio of chip/heatsink height to board gap to calculate a roughness number for these calculations. Figures 6 and 7 show the equilibrium air flow volume and velocity, respectively, versus pressure in the card cage; the PXFL cards have 16.5mm (0.65) board gap.

Heat sinks are required on all chips with power dissipation of 4 watts or higher (some 45 chips on each system slot). The size and design of the heatsink for a given chip were determined, based on the worst-case position on the board, to keep junction temperatures at 100^oC or lower. We chose to use only two different kinds of packages for the chips on the Render board, and to use the same type of heatsink for all chips, thus reducing assembly errors and manufacturing costs for the boards. The heatsinks for the Render board are all 6.6mm tall, and are either 28mm or 40mm square. The high-dissipation chips are mounted in MQUAD (metal quad flatpack), cavity-down, packages, chosen for their high pin count (up to 304 pins for one chip type), good electrical characteristics for high-speed chip I/O circuitry, and low θ_{jc}.

In single-sided surface-mount designs, the PC board itself can carry away much of the heat from MQFP package leads to the cooling air on the backside of the board. Our board design, however, has identical chips mounted in mirror image on both sides of the board at the same location, so heat must be removed entirely from the top of the package, aided by heatsinks. Thermal performance of MQFPs in this situation is not available in the literature, so we computed new θ_{ja} equations to compensate for lack of through-board cooling. Flow analysis of the board in both 2D and 3D showed that, even with 5 m/sec of airflow, chip packages at the bottom (exhaust side) of the board could not be maintained at junction temperatures below 100^oC without heatsinks. The sinks we are using are unidirectional, with fins chopped to provide a localized turbulent environment. Overall airflow between boards remains laminar, to reduce problems of air eddies.

5. EMI Filtering

Electro-magnetic shielding for the 200Mhz computer is divided into the two standard areas: conducted and radiated emissions. Conducted emissions are first dealt with at the 48 VDC switching front end power supplies. The supplies are power factor corrected to assist in meeting European EMI standards. In addition to their internal filtering, a large filter is placed right after the line input to filter out any additional harmonics produced from the ancillary

rack mounted modules that could be plugged into the internal power distribution network for custom or user configurations.

Radiated emission shielding is accomplished through three layers of shielding and design. The first layer, at the board level, is the most important. Clocks on the board are set at the lowest common frequency (100MHz) and are distributed as sinusoidal signal, thereby eliminating harmonics. Board-to-board signaling uses controlled-risetime, very low voltage signaling carried entirely in tri-plate (fully shielded) transmission lines. The board also has two outer layers of ground coverage to further assist in reducing radiated EMI from internal signal lines. Fully shielded connectors are used between daughter cards and backplane, both to maintain signal integrity in the low-voltage signaling paths and to reduce EMI. The second layer of shielding is at the card cage level, including metal card front panels and wire-mesh shields at the top and bottom of the card cage. A full metal system enclosure forms the third layer of EMI shielding.

Conclusion

In very high density situations, manufacturer's thermal data for chip carriers is normally too optimistic. However, FEA fluid flow analysis combined with detailed analytical studies can characterize commercial chip carriers to be used in high density / high power systems where often exotic and expensive cooling methods need to be employed. Using the same detailed analysis for system wide design will also reduce costs in such ways as splitting large system boards into smaller, cheaper boards while distributing the expense of connectors by sharing pins and backplane real estate by having boards plug into both sides of a backplane.

Acknowledgments

This research is supported in part by the Advanced Projects Research Agency and the National Science Foundation. Special thanks to all who have contributed to the Pixel-Planes and PixelFlow projects, the Microelectronic Systems Laboratory and to thank Hewlett-Packard Corporation for their support.

References

Alexander, Thomas B., Kenneth G. Robertson, "Corporate Business Servers: An Alternative to Mainframes for Business Computing," *Hewlett-Packard Journal*, June 1994 pp. 25 - 30.

Cramblitt, Bob, "Worlds Fastest Graphics Computer," *Computer Graphics World*, March 1993, pg. 19.

Dash, Glen A., Editor, *Compliance Engineering, 1993 Reference Guide*, Boxboro, MA, Compliance Engineering, 1993.

Fuchs, H., J. Poulton J. Eyles, T. Greer, J. Goldfeather, D. Ellsworth, S. Molnar, and L. Israel (1989). "Pixel-Planes 5: A Heterogeneous Multiprocessor Graphic System Using Processor Enhanced Memories," *Computer Graphics* (Proceedings of SIGGRAPH '89), pp. 79-88.

Mardiguian, Michael, *Controlling Radiated Emissions by Design*, New York, Van Nostrand Reinhold, 1992.

Molnar, Steven, John Eyles, and John Poulton, "PixelFlow: High-Speed Rendering Using Image Composition," *Computer Graphics* (Proceedings of SIGGRAPH '92), pp. 231 - 240.

Ozisik, Necati M., *Heat Transfer a Basic Approach*, McGraw-Hill, Inc. 1985.

Tummala, Rao R. and Eugene J. Rymaszewski, Editors, *Microelectronics Packaging Handbook*, New York, Van Nostrand Reinhold, 1989.

Yamaji, Y., Y. Atsumi and Y. Hiruta, "Thermal Characterization of LSI Packages Mounted on PC Boards: Evaluation of the Thermal Effects of PC Boards," *Advances in Electronic Packaging* (Proceedings of the 1992 ASME / JSME Conference on Electronic Packaging), pp. 199 - 205.

"Z-PACK 2mm HM Interconnection System," AMP specifications book, November, 1993.