Cooling High-Power Dissipating Artificial Intelligence Chips Using Refrigerant

By: Waheeb Mukatash, Tyler Yang, Matthew Moscoso, Marc Massalt, Merari Mejia Robles, Charlie Nino

Scope

We are trying to address the need for cooling high-power dissipating chips, particularly AI chips. These chips consume more power than previous chips. Thus, we intend to design, build, and test a vapor compression refrigeration cycle to analyze its efficiency in cooling 4 AI chips. We will compare our prototype to previous research and provide steps to follow to continue improving refrigeration for high-power dissipating chips.

Problem Statement

As developers transition to a realm of artificial intelligence, these high-performing circuits will expose issues with existing cooling systems since they will not surmount the needed output and productivity. Refrigeration cooling can directly provide solutions to pressing issues such as limited power envelopes, high upfront expenses, system inefficiencies, as well as industry sustainability. Current trends towards high performance computing shows that the power consumption per chips and rack units are requiring more energy which will naturally equate to more heat being generated and thus, more cooling being demanded.

More data is becoming readily available and must be processed on demand efficiently and effectively. With flagship products such as NVIDIA's H100 supercomputing node requiring upwards of 700W of power to reach maximum performance reveals the dire need to move toward another cooling approach. In terms of performance, air cooling possesses the capabilities

of only regulating 30 kW per rack with major inefficiencies [1]. With refrigeration cooling being able to triple that capability by being able to cool 100kW per rack, the use of AI chips real-time processing becomes possible to achieve smarter and quick computations through practically unlimited cooling capabilities that the dielectric fluid provides. These future overworked processors thus require a significant boost in cooling power and must be geared toward decreasing leakage as well as evolving the power performance.

Introduction/Background

As of today, conditioning systems such as conventional air cooling and liquid cooling are the primary systems used to regulate temperatures within high-powered chips. Computing chips have evolved over time, requiring more power dissipation. In 2008, IBM introduced the Power 575 Supercomputing Node which offered a fully configured cooling processor that contained the potential of dissipating 72 kW via water cooling. Most liquid cooling systems today mirror IBMs processes dating even further than two decades back prompting a new revolutionary system. NVIDIA's state-of-the-art and most commonly used GPU, H100, is used specifically for Artificial Intelligence, AI, and includes 80 billion transistors. This equates to a 160% power consumption increase from the previous generation's chip [2]. Current cooling configurations would require compatibility and optimization changes to effectively manage these unprecedented heat loads. Our project is geared to focus on researching using a refrigeration cooling system to cool high-power dissipating AI chips.

Technology's next advancement, artificial intelligence, projects the future direction of how engineers interact with automated engineering. Artificial intelligence is the basis of computer algorithms to process information and mimic human intelligence which started in the

early 1950s and led to machine learning technologies of today [3]. Some applications include autonomous vehicles, voice assistants, automated predictions, modeling, facial recognition, and many other useful applications. Over the years, the development of AI core chips has experienced many changes as they originally started as a simple semiconductor to be able to handle large amounts of data and be power efficient. Compared to traditional general-purpose CPUs and chips, AI chips require much more computing power, speed, and efficiency. Today's AI chips include graphics processing units (GPUs), an application-specific integrated circuit, and field-programmable gate arrays (FPGAs) [4]. The newer transistors showcase substantial improvement in running speed and energy efficiency compared to traditional large transistors. Traditional CPUs lack the processing performance needed while GPUs are capable of handling parallel processing and can be used in basic AI applications to enhance neural networks. AI hardware generates an excessive amount of heat, and thus thermal management is key to creating a cooling system for these high-powered dissipating chips. [5]

The vapor compression evaporation cycle historically repels from electronics because of the many challenges that could arise, like leakage and connection length. However, it is "one the most promising alternative cooling techniques for high heat dissipation electronics cooling." [6] The vapor compression evaporation system consists of four main components: the evaporator, the compressor, the condenser, and the expansion valve. The refrigerant leaves the evaporator as a slightly superheated vapor and enters the compressor. If this path is long, the pressure drops as a result of fluid friction and heat transfer from the surroundings. As a result, the specific volume increases causing the compressor to require a higher power input to operate at a steady-flow rate. The compressor causes the pressure to increase and thus, increases the temperature. The

high-pressure, high-temperature refrigerant then enters the condenser to convert the liquid to a subcooled liquid. This improves cooling capacity and prevents vapor bubbles from entering the expansion valve. The expansion valve's throttling effect separates the high and low-pressure sides of the cycle and decreases saturation temperature. Thus, the refrigerant will boil at a low temperature in the evaporator. The connecting line between the expansion valve and the evaporator must be short to avoid decreased pressure [7]. The space allotted in electronics becomes a challenge when designing cooling systems for chips. Developing a vapor compression evaporation cycle that is efficient and compact enough to fit in a limited space is an ongoing challenge [6]. In addition to the bulkiness of the system, liquid and refrigerant systems have leakage problems [8]. Further work has to be done to integrate liquid or refrigerant cycles into semiconductors [9]. Our project implements a vapor compression evaporation system utilizing Refrigerant 134A's and proper housing unit, power requirements, and arrangements of components.

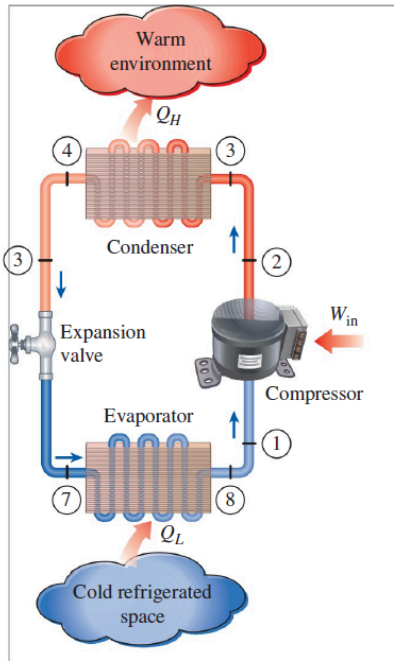

Diagram of the Vapor-Compression Evaporation Cycle [7]

Literature Survey

There have been plenty of studies and research done to reduce the amount of heat, energy, and power produced by high-powered chips. With the rise in artificial intelligence (AI), there will be an even bigger demand for a new thermal design due to the power generated by AI chips [2]. In analyzing cooling methods to meet these thermal demands, different cooling methods must be analyzed. Purdue [6] references that Scott [10] classifies electronic cooling techniques, using refrigeration, into four categories: using refrigerants to cool liquid or air, refrigerating heatsinks, using liquid nitrogen baths, and thermoelectric cooling. In comparing the refrigeration of air or liquid technique to the refrigerated heatsink method, Purdue's [6] general findings are that using a refrigerated heatsink has a lower temperature at the cold surface. In addition, because the evaporator would be mounted directly on the electronic component, the electronic temperature is colder than the surrounding temperature and the refrigerated heatsink method is more compact than using a cooling fluid. For liquid nitrogen baths, immersion would require a dewar flask - a container made of insulated material. Special care is required for this method, to achieve advanced insulation - greater risk of leakage. Lastly, the capacity and overall effectiveness of the thermoelectric cooling method are minimal and not promising [6]. The first IBM system to employ refrigeration cooling techniques is the IBM S/390 G4 CMOS server system. The processing module is the element cooled by refrigeration; its form factor is $267\times267\times711mm^3$ and has a weight of 27 kg. To absorb moisture the evaporator contains 260 grams of silica gel. For this G4 server, the average process module temperature was about 40℃ which is 35℃ lower than the air cooling system of the same design. These IBM server units have been tested and implemented to ensure high reliability [11].

A similar experiment to the one proposed was done by Rajiv Mongia *et al.* [12], a VCRS was built to cool down the CPU of a Notebook computer. The working fluid used for this experiment was Isobutane, also known as R600a. Their refrigeration system consisted of the following components: An evaporator, a condenser, a throttling device in the form of a capillary tube, and a compressor. In addition to the components to make up the refrigeration cycle, sensors were attached at several points throughout the system to measure the temperature and pressure of each component. The data was recorded using a computerized DAQ system that was programmed with LabVIEW™. Several experiments were performed, each time making modifications to either a component or an operating condition. For example, for some experiments, the airflow at the condenser was increased or the evaporator (DTTV) power was decreased. The results of their experiment were evaluated through the coefficient of performance (COP). They were able to achieve a COP of 2.25 and greater depending on the variables changed. The highest COP achieved through their experimentation was 3.70 which occurred when the DTTV power was reduced. A slight increase in COP was also observed when the airflow at the condenser was increased. Comparing the measured COP and the ideal COP (Carnot refrigeration system), they reached about 25-30% of the Carnot efficiency, which increased as the pressure ratio increased. They found that a higher pressure ratio could be achieved through having a higher gas temperature.

Another solution was evaluated by Lung-Yue Jeng and Tun-Ping Teng [8]; a hybrid design of a liquid cooling system and a vapor compression refrigeration system (VCRS). To achieve maximum efficiency a combination of aluminum oxide nanoparticles and water was used because aluminum oxides disperse easily in water and have good suspension performance.

The liquid coolant was operating at a flow rate of 2.0 L/min. Hydrocarbon refrigerant was used for the VCRS as it improves the coefficient of performance, cooling capacity, and is environmentally friendly. To simulate the power of a CPU a heating module made up of "pure copper blocks were placed in four tubular heaters (150 W x 4DC100V) and placed alongside a heat exchanger with thermal paste." Two power supplies were connected to the tubular heaters to simulate the heat generation. The components that were used for this thermal design were expansion valve, control valve compressor, pump, condenser, flow meters, and air-cooled heat exchanger. When compared to the other test methods performed, which include liquid cooling with distilled water, aluminum oxide, and using VCRS with different types of hydrocarbon coolants, the hybrid design proved to be the best. It was able to have a cooling capacity of 540 W, a surface temperature of the heater at 75.8℃, and a power consumption of the compressor and water pump at 29.5 W. The hybrid design also proved to have a cooling capacity 4.5 times higher than the heat of the CPU.

Some of the changes going to be performed are using an AI chip, R134a refrigerant, and using a VCRS. . The size of each component will be greater compared to the study done by Rajic Mongia *et al.*, as the proposed experiment does not hold the same form requirements as the studied experiment, partially due to the increased power dissipation requirements. The proposed experiment will also use an expansion valve rather than a capillary tube. Additionally, the working fluid for this system will be changed to R134a rather than the R600a used. To add to this the coolant R134a has proved to be an efficient coolant when used in spray cooling and is considered to be environmentally friendly [13]. We look to implement a vapor compression refrigeration system as it can be the most suitable for AI chip applications. Its benefits include a

low mass flow rate, a high COP, and the ability to transfer heat away from the source with low cold plate temperatures [6]. To expand upon this refrigeration technique we look to research, explore, and build a working prototype that uses a refrigerated cooling method in which a heat sink would operate below ambient temperature - allowing for optimal cooling of high power dissipating AI chips.

Methodology

After completing the preliminary research and literature review, we will begin designing a physical model of a system that uses VCRS cooling. This system will include all the major components of a vapor compression refrigeration system, a cold plate that will be mounted directly on the component being cooled, as well as a heat exchanger. We intend to build a prototype that will house the 8 Thermal Test Vehicle (TTV) chips that have identical properties to that of 8 AI NVIDIA chips, dissipating a power of ~1.1kW each. These TTV chips will be connected to an external power source to model the heat generated. The end goal will be to have a working prototype of the system that models the capabilities of refrigeration cooling. We will use computer-aided design (CAD) software to model our design. We will perform CFD analysis for thermal and airflow numbers and calculate the system's overall efficiency and thermal properties before physically prototyping. We will seek industry knowledge and sponsors to gather components and gather the necessary funding for our prototype. Once we are able to build our prototype, we will perform testing on our model. This testing will be done with a DAQ system that is connected to sensors located throughout various parts of the VCRS: before and after the compressor, the evaporator (TTV), the condenser, and the expansion valve. We will make use of transducers and thermocouples to get the pressure and temperature at each location. This will allow us to determine the heat dissipation of the refrigerant as well as verify that each

component of the system behaves as theorized. The testing will be done at multiple levels of power dissipation leading up to the power dissipation of the 8 AI NVIDIA chips. The COP will be determined at each of these levels to determine the effects of high heat dissipation on the system.

Timeline

TASK NUMBER	TASK TITLE	TASK OWNER	START DATE	DUE DATE	DURATION (DAYS)
1	**Proposal**				
1.1	Introduction	Charlie, Merari Tyler	9/13/23	9/20/23	7
1.1.1	Literature Review	Marc, Matthew, Waheeb	9/13/23	9/20/23	7
1.2	Scope	Merari	9/20/23	9/27/23	7
1.3	Problem Statement	Tyler	9/20/23	9/27/23	7
1.4	Methodology	Marc and Waheeb	9/20/23	9/27/23	7
1.5	Timeline	Matthew	9/20/23	9/27/23	7
1.6	Outcome	Charlie	9/20/23	9/27/23	7
1.7	Presentation #1	Everyone	9/13/23	10/4/23	21
2	**Project Oral Presentation #2**				
2.1	Initial Design	Marc, Matthew, Merari, Waheeb	10/4/23	10/18/23	14
2.2	Calculations	Waheeb	10/18/23	10/25/23	7
2.2.1	Engineering Analysis	Waheeb	10/18/23	10/25/23	7
2.3	Presentation #2	Everyone	10/4/23	11/1/23	27

3	**Project Oral Presentation #3**				
3.1	Material Selection	Marc	11/1/23	11/15/23	14
3.2	Part Specification and Selection	Marc, Matthew, Merari, Waheeb	11/1/23	11/15/23	14
3.3	Purchase Parts	Charlie and Tyler	11/8/23	11/15/23	7
3.4	Cost Analysis	Charlie and Tyler	11/8/23	11/15/23	7
3.5	Final Design	Marc, Matthew, Merari, Waheeb	11/15/23	11/22/23	7
3.6	BOM	Charlie and Tyler	11/15/23	11/22/23	7
3.7	Fabrication Plan	Charlie and Tyler	11/15/23	11/22/23	7
3.8	Presentation #3	Everyone	11/1/23	11/29/23	28
4	**Final Report**				
4.1	Final Report	Everyone	11/29/23	12/6/23	7
4.2	Individual Evaluation	Everyone	11/29/23	12/6/23	7

Expected Outcome

We expect that refrigeration cooling will be a more effective cooling method than water cooling for these AI electronic applications. We expect this due to the advantage of operating at a sub ambient temperature, something water cooling is incapable of. Possible further areas of research would be to explore various types of refrigerants and compare them against each other. In addition more testing can be done in a different form factor as well as further investigating different types of heat exchangers and heat sinks.

Works Cited

[1] Adminxs. (2022, January 31). *Is immersion cooling the future of high-performance computing?*. Green Revolution Cooling. https://www.grcooling.com/blog/immersion-cooling-future-performance-computing/

[2] Ellsworth, M. J., Jr., Goth, G. F., Zoodsma, R. J., Arvelo, A., Campbell, L. A., and Anderl, W. J. (June 11, 2012). "An Overview of the IBM Power 775 Supercomputer Water Cooling System." ASME. J. Electron. Package. June 2012; 134(2): 020906. https://doi.org/10.1115/1.4006140

[3] Jotrin. (2022, January 4). *A brief history of the development of Ai Chips*. Jotrin Electronics. https://www.jotrin.com/technology/details/a-brief-history-of-the-development-of-ai-chips

[4] Saif M. Khan and Alexander Mann, "AI Chips: What They Are and Why They Matter" (Center for Security and Emerging Technology, April 2020), cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter/. https://doi.org/10.51593/20190014 https://www.grcooling.com/blog/immersion-cooling-future-performance-computing/https://www.grcooling.com/blog/immersion-cooling-future-performance-computing/

[5] *What is AI chip design? – how it works*. Synopsys. (2023). https://www.synopsys.com/ai/what-is-ai-chip-design.html#:~:text=Traditional%20CPUs%20typically%20lack%20the,be%20applied%20to%20AI%20applications.

[6] Trutassanawin, S., & Groll, E. A. (2004, July). Review of refrigeration technologies for high heat dissipation. Purdue e-Pubs. https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1677&context=iracc

[7] Çengel, Yunus A., et al. Thermodynamics: An Engineering Approach. McGraw-Hill Education, 2019.

[8] Jeng, L.-Y., & Teng, T.-P. (2013). Performance evaluation of a hybrid cooling system for electronic chips. Experimental Thermal and Fluid Science, 45, 155–162.

[9] Blinov, A., Vinnikov, D., & Lehtla, T. (2011). Cooling methods for high-power electronic systems. Scientific Journal of Riga Technical University. Power and Electrical Engineering, 29(1), 79–86. https://doi.org/10.2478/v10144-011-0014-x

https://doi.org/10.1016/j.expthermflusci.2012.10.020

[10] Scott, A.W., 1974, Cooling of Electronic Equipment, John Wiley and Sons, pp. 204-227.

[11] Schmidt, R.R., and Notohardjono, B.D., 2002, High-End Server Low-Temperature Cooling, IBM Journal Research and Development, Vol. 46, No. 6, November, pp. 739-751.

[12] Mongia, R., Masahiro, K., DiStefano, E., Barry, J., Chen, W., Izenson, M., Possamai, F., Zimmermann, A., & Mochizuki, M. (2006). Small scale refrigeration system for electronics cooling within a notebook computer. *Thermal and Thermomechanical Proceedings 10th Intersociety Conference on Phenomena in Electronics Systems, 2006. ITHERM 2006.* https://doi.org/10.1109/itherm.2006.1645421

[13] Groulx, D., & Kheirabadi C. A. (2016). Cooling of server electronics: A design review of existing technology. *Applied Thermal Engineering,* 105, 622-638. https://doi.org/10.1016/j.applthermaleng.2016.03.056.