SUMMARY

The Customers

Cooling optimization—cooling cost minimization without increasing risk of service interruption and performance degradation—poses a critical problem for the information technology (IT) industry. Cooling systems (e.g. computer room air conditioning (CRAC) unit), dedicated to waste heat removal to ensure service availability and reliability, account for nearly 40% of data center electricity consumption, which has risen to 2% of world electricity usage. Workload-proportional cooling (i.e., cooling in-sync with changes in IT workload), which could lead to significant savings in cooling costs, is yet to be implemented. Due to complexity of the problem, the related data center facility management (DCFM) industry highly fragmented in nature with no entity providing a holistic energy management solution. AdeptDC co. is proposing the development of a real-time software appliance (SA) that would automatically adjust cooling set-points based on IT device temperatures. The proposed automated cooling technology would promote workload-proportional cooling and lead to cooling optimization. AdeptDC's primary customer segments will be the managers of data centers (e.g. Bloomberg), and data center design firms (e.g. DLB Associates). The potential channel partners would be building management system (BMS) vendors (e.g. Johnson Controls), component vendors (e.g. Intel), and original equipment manufacturer (e.g. Emerson-Liebert). In absence of a reliable workload-proportional cooling technology, the customers currently use wasteful peak-provisioned cooling. In contrast, AdeptDC's SA will provide a real-time predictive analytics framework that would enable workload proportional cooling.

Value Propositions

The key value proposition is reduction of data center cooling cost without increasing risk of service interruption and performance degradation. While current cooling solutions for data centers provide wasteful peak-provisioned cooling, the proposed SA would provide following benefits:

- Cooling cost savings by 15-25%
- Cooling capacity improvement by 20-30%
- Automated high-precision optimal cooling control

Key Differentiators

Cooling Factor	State-of-the-art	Proposed Technology	Advantages
Control Variable	Equipment inlet air temperature	IT device temperature	Efficient control
Allocation Algorithm	Static, Reactive	Real-time, Predictive	Adaptive control
Optimization	Single equipment	All cooling systems	Holistic control
Automation	Does not exist	Automated	Precise control

Innovation

AdeptDC's SA will enable workload-proportional cooling in data centers. First, it will use IT device (e.g. CPU) temperature as the control variable unlike the state-of-the-art (e.g. Panduit) solutions which use equipment inlet air temperature. This would make AdeptDC's innovation better in cooling need assessment and control. Second, the proposed SA will use a real-time predictive algorithm that computes optimal cooling set-points to enable workload-proportional cooling. The AdeptDC's SA therefore avoids resource waste resulting from peak-provisioned cooling currently provided by the state-of-the-art solutions. Third, the state-of-the-art solutions, which compute optimal cooling set-points for each equipment separately, could be counter-productive. For example, the energy saving from a CRAC unit operating at a high temperature might work against the energy loss from the server fans which need to operate at a high temperature environment. The proposed SA will compute optimal cooling set-points for all cooling systems to promote holistic energy optimization. Finally, the proposed technology will link the computed optimal set-points to the BMS and will automate optimal cooling. Such an automated optimal cooling solution that integrates IT and cooling systems in data center does not exist in the DCFM industry.

MARKET OPPORTUNITY

Data centers are the epicenter of the digital world. They host the servers, network switches, and storage disks for computing, transmitting, and storing digital information. Major IT-service providers and e-commerce companies (e.g. Google, Amazon) have their own enterprise data centers. Other small and medium-sized organizations depend on commercial data center hosting companies such as Equinix to provide them with data center services. The data center hosting industry has undergone rapid growth from $28 billion in 2009 to $60 billion in 2013. Operating a data center requires a tremendous amount of electricity (at the rate of few to tens of MW-s) to feed its IT equipment and its facility infrastructure (e.g. cooling systems like CRACs), driving energy consumption to 2% of the worldwide electricity usage. In 2013, data centers in the United States consumed nearly 91 billion kWh of electricity, at a cost of approximately $7.5 billion . By 2020, data center operation will consume 140 billion kWh of electricity and cause 150 MMT of carbon pollution annually. With soaring electricity costs and enhanced environmental concerns, market literature suggests that data center energy efficiency products have significant growth potential. Pike Research reports that the worldwide market for green data centers would grow at a compound annual growth rate (CAGR) of almost 28% from $17.1 billion in 2012 to $45.4 billion by 2016. Since data center cooling uses approximately 40% of the electricity intake, cooling optimization is a major problem in improving data center energy efficiency. This is accompanied by lack of coordination between IT and cooling management systems, which is complicated both technically and politically. With the proposed workload-proportional cooling optimization SA, AdeptDC will primarily operate in a segment of data center industry called data center facilities management (DCFM) which is a $1.3 billion market with CAGR of 39%. DCFM industry encompasses the tools that monitor, manage, and control energy consumption of data center facility infrastructure components. These tools are primarily used by data center facility managers. After its inception in 2009 mostly with data center monitoring systems [6], DCFM industry is currently entering into an early growth phase as data centers are beginning to adopt DCFM tools with clearer ROI projections. At present, the DCFM industry is ripe for a value-added disruptive innovation. In terms of technical sophistication, DCFM tools can be divided into five different maturity levels . Depending on their present market penetrations, these levels can be further segmented into three categories:

100% Market Penetration

- Level 1 (Basic): These tools (e.g. Nform) enable basic equipment-level monitoring
- Level 2 (Reactive): These tools (e.g. SiteScan web) are useful for software-based monitoring of environment and equipment power usage, change management, and asset planning in data centers.

60% Market Penetration

- Level 3 (Proactive): These tools (e.g. Trellis) are useful for efficient infrastructure management, resource provisioning, utilization tracking, and risk mitigation.

30% or Less Market Penetration

- Level 4 (Optimizing): These tools (e.g. Vigilent) are useful for predictive maintenance, service management, and near real-time resource optimization.
- Level 5 (Autonomic): Still at the embryonic stage, these tools provide software-based data center control to regulate data centers according to some pre-assigned rules and service-level agreements. AdeptDC's SA will be in this level.

The market penetration data clearly suggests the potential growth opportunities for AdeptDC's SA as a Level 5 DCFM tool. The key economic and market drivers for the DCFM industry are:

1. Resource Optimization: With denser integrated circuits and virtualization-based server consolidation technology; heat load on data center cooling system is escalating. Increasing power consumption and greenhouse gas emissions have warranted the need for resource-optimized data center operation. Saving energy cost by increasing coordination between IT and cooling systems is likely to get significant market traction. Sourcing skilled personnel is another acute problem faced by data center. Therefore, automated data center management solutions will be widely adopted.
2. Data-driven Predictive Analytics: Most IT devices are equipped with on-board temperature monitoring sensors. The DCFM industry is looking for solutions that can use these sensor data and convert them into actionable information. Real-time predictive analytics [3] can draw useful insights from these data and is likely to be used as an on-demand resource allocation mechanism.
3. Reliability: The primary goal of data center cooling is to keep the IT device temperature below a reliability limit to guarantee uninterrupted computing services. Without a workload-proportional control technology, the facility managers often over-provision cooling to ensure service reliability.

Market Validation

Last July, some of the executive board members of AdeptDC (Rajat, Mark, Yogendra) participated in the NSF I-Corps program and interviewed 112 potential customers. The useful insights gained during these customer interviews helped AdeptDC to consolidate its value proposition, understand the pertinent commercial landscape, and identify potential product-market fit. Moreover, attendance at various data center tradeshows (e.g. 7x24 Exchange), and review of market literature (e.g. Gartner) have assisted AdeptDC to stay abreast of the latest market trend. Critical industry acclaims received on one of the founder's (Rajat) technical blog that introduces the concept of AdeptDC's software appliance further attests the market need.

Business Model

AdeptDC co. will develop and market a software appliance (SA) for workload-proportional cooling of data centers. The revenue would be derived from yearly subscription fee which would include software license, maintenance, hardware, and implementation. Another important revenue stream could be royalty and commissions per sale collected from channel partners like Intel and Johnson Controls. For an average (10 MW-rated) customer, the yearly revenue estimate is $140K (Table 1).

Table 1: Yearly subscription-based revenue model from one average (10 MW-rated data center) customer where cost-savings to the customer would approach $1 million per year

License Fee	$100,000
Maintenance Fee	$10,000
Hardware	$5,000
Implementation	$25,000
Total Yearly Revenue	**$140,000**

A laboratory-scale prototype shows that proposed SA can save 10-30% of data center cooling cost and improve cooling equipment capacity by 20-40%. For a 10 MW data center, that is equivalent to $1-$1.2 million cooling cost savings, and capacity utilization improvement amounts to $1.3-$1.5 million per year.

Competitive Landscape

DCFM is closely related to the data center infrastructure management (DCIM) industry. A recent Gartner report identified 18 prominent players in the DCIM/ DCFM industry. Based on the company vision and the execution capability, these companies can be divided into four categories:

- Leaders: Emerson Network Power, Schneider Electric, CA Technologies, Nlyte Software
- Visionaries: Panduit, Cormant, FNT, IO
- Challengers: Raritan, iTRACS
- Niche Players: FieldView Solutions, ABB, Optimum Path, Modius, Device42, Geist, Rackwise

In keeping with the segmented nature of the market, no company is capable of developing a workload-proportional DCFM tool that can provide an integrated IT and facility management capability. The reason is rooted in the operational silos currently in data center IT and facility components. It is difficult for a single company to own cross-functional capabilities in IT and cooling management. IT-centric companies like CA Technologies offer competent IT management tools, but lacks proper facility management capability. They are strongly dependent on the non-exclusive partnerships with other companies (e.g. CA Technologies partners with Optimum Path). Facility-oriented companies like Emerson, Schneider, and Panduit, on the other hand, depend on other IT management companies for putting together a comprehensive infrastructure management suite. Therefore, the industry is characterized by numerous merger and acquisition activities and partnership agreements. A few examples are: Emerson's acquisition of Aperture and Avocent, Schneider Electric's partnership with Vigilent, and Panduit's acquisition of Synapsense. In general, big players are using their business resources to consolidate their technical weaknesses and gain competitive advantages. Niche players like Optimum Path, in contrast, depend on the partnerships with the leaders such as CA Technologies for their survival. Therefore, niche players strongly focus on innovation to attract big players. Big players, on the other hand, focus strongly on acquiring innovative small companies, and strategic spending of their marketing dollars. Niche players like Rackwise are finding it difficult to survive alone even though they have a good technical solution. AdeptDC is expecting to enter the DCFM market in the first quarter of 2016. By then, the DCFM industry might have a few tools to enable automated workload-proportional cooling. The competition will be very intense in the DCFM market. At this moment, the area is still under research and development. But, by the end of 2015 this will likely be an intense focus of the market.

Market Risk

Following are the key risks in bringing the innovation to market:

1. Getting access to IT data: Two groups in a traditional data center organization, namely IT and facilities groups, do not work well with each other. Although AdeptDC's SA will be primarily used by the facility managers, it will need access to IT temperature data. To overcome this potential problem, the proposed SA will provide an IT temperature (e.g. CPU temperature) data aggregation tool which will be easy to deploy and intuitive to use. Another side of the problem is IT access control: at least, 22% of the data centers are multi-tenant hosting service providers. Approximately 40% of these data centers do not have access to critical data such as CPU temperatures. An alternative viable strategy would be use server electrical power data as the control basis. With data-driven control algorithm, the proposed software appliance can flexibly handle different data types.

2. Awareness: According to a survey , 86% of the DCIM industry seeks a better cooling optimization tools. However, risk-averse facility management personnel might oppose AdeptDC's SA due to its real-time and autonomous nature. This demands technical high-fidelity feasibility studies in some alpha customer sites such as Georgia Tech Office of Information Technology (OIT) Data Center which handles network traffic from Georgia Tech, Emory, and UGA. In addition, regular tradeshow and symposium visits will be directed to educate customers. Moreover, the proposed software appliance will have a user-interactive dashboard where the user can easily choose to shift to the manual control, when needed.
3. Partnership: As discussed before, DCIM/ DCFM industry is highly segmented in nature. An innovative startup will enter into the niche player category. It will need to develop a partnership with a bigger player(s) to survive. That involves various factors such as overall business climate of the industry and strategic intellectual property acquisition.

External Partners

AdeptDC needs few partners to bring the proposed innovation to market. The key resources needed are:

1. Testing Facility: For technology validation, the access to a high-fidelity test facility is critically important. This facility should provide tightly controlled environments where the proposed SA can be deployed and tested for various cases. Good candidates for the test facilities are Georgia Tech OIT Data Center and/or Department of Energy (DOE)/ Lawrence Berkley National Laboratory Data Center (LBNL). For SBIR Phase-I research, the testing facility will be needed in October, 2015. The letters of support from Georgia OIT and LBNL are attached with this proposal.
2. Distribution: Given the risk-averse nature of the DCFM industry, it is strategically critical to partner with other big players. Smaller DCFM players like Vigilent often adopt a similar strategy when they partner with big companies like Schneider Electric. Various other firms in the data center industry could be our potential channel partners, such as: contractor firms like Batchelor & Kimbell, manufacturers' reps like Joe Powell Associates, original equipment vendors (OEM-s) like Liebert, and design firms like DLB Associates. This distribution assistance will be needed during the latter part of the research, tentatively sometime around November, 2015. The letter of support from DLB Associates is attached with this proposal.
3. Technical Implementation: The effectiveness of AdeptDC's SA demands its seamless integration and coordination with other hardware. It dictates I/O versatility to handle various types of communication protocols and smooth integration with other control units such as building management systems (BMS-s) and programmable logic controllers (PLC-s). In that regard, BMS vendors like Johnson Controls—leading provider of BMS—could be strategic partners. For this project, the technical implementation support will be needed during the benchmarking study, tentatively sometime around October, 2015.

THE INNOVATION

Cooling optimization—minimization of cooling cost without increasing the risk of service interruption and performance degradation—is a critical data center management problem because of its huge cost saving potential and environmental impact. To solve this problem, AdeptDC is proposing to develop a SA that will promote workload-proportional cooling in data centers. The preliminary technical and market research activities have led AdeptDC to the hypothesis that the cooling optimization problem can be

solved by automating cooling in sync with IT workload variation. The proposed SA will have three fundamental capabilities: first, it will monitor IT device temperatures (e.g. CPU temperatures) and archive them into a training database. Second, it will use a machine learning-based real-time predictive analytics framework to compute workload-proportional optimal cooling set-points. Finally, the computed optimal set-points will be linked to the BMS to control corresponding cooling hardware components. AdeptDC has already developed a simulation-based laboratory-scale prototype of the technology and estimated its cost-saving benefits.

AdeptDC will adopt a proprietary data-driven predictive analytics framework, which enables logarithmic time computation. As a result, the framework is useful as a real-time control algorithm. The framework will use IT device temperatures for different types of IT workloads compiled over a range of cooling operating points as the training data. The IT device temperature data will be collected via a multi-protocol-based network management system deployed in a data center. The real-time predictive analytics framework will predict optimal cooling set-points that ensure cost-efficient cooling. The optimal cooling allocation will start by determining CPU temperature data for the most cost-efficient cooling points, and assessing whether the reliability constraints are satisfied. If they are not satisfied, an iterative procedure that transfers to the next cost-effective cooling set-point is invoked. Given the real-time nature of the underlying algorithm, overall iterative procedure takes only a few (<10 s) seconds. This ensures cooling allocation is in-sync with the dynamic IT workload (e.g. transient CPU temperature/ CPU utilization or other representative variable like rack current draw). Finally, the workload-proportional set-points will be transferred to the various cooling units to implement optimal cooling in data centers. Depending on data center cooling topology and customer preference, this transfer will take place through the building management systems (BMSs) or other localized control units (e.g. PLCs) using different communication protocols like Modbus or BACnet.

To estimate the energy savings potential of the proposed technology, a simulation-based case study has been conducted in the CEETHERM Data Center Lab at Georgia Tech. In this particular case-study, two cooling units were used: one computer room air conditioning (CRAC) unit (rating~150 kW) and one rear door heat exchanger (RDHx) unit (rating~18 kW). As compiled in Table 2, the average energy savings figures in these two cooling units are assessed for four different types of workloads—cloud, batch, enterprise, and high-performance (HPC). The saving percentages are computed based on an industry-standard cooling baseline defined as 18 ^{0}C CRAC supply temperature and 12 psi RDHx operating pressure. The cost modeling for CRAC unit is largely based on the whitepaper in . It is assumed that the operating cost for RDHx is directly proportional to the square of the operating pressure. The reliability constraint in this case study is defined as the CPU temperature is less than or equal to 65 ^{0}C.

Table 2: Savings potential benchmarking against 18 ^{0}C CRAC supply temperature and 12 psi RDHx operating pressure.

Workload Type	**CRAC (150 kW) Saving**	**RDHx (18 kW) Saving**
Cloud	19%	50%
Batch	17%	63%
Enterprise	20%	50%
HPC	4%	41%

Technical Risk

The key technical challenges in bringing the innovation to market are:

1. IT Data Access: Getting access to IT data is difficult for technical and political reasons. Often, data center facility managers are not trained enough to extract IT temperature data. In wholesale colocation data centers (~19%), customers manage IT equipment. Therefore, getting access to IT data for the facility management often nearly impractical. Deploying an IT data extraction system requires working around data center network traffic that poses an additional level of complexity.

2. Real-time Analytics: The pivotal feature of the proposed SA is a logarithmic time control algorithm. However, its real-time implementation demands efficient coding and network deployment. While the programming efficiency is dictated by the accuracy specification by the users, the network deployment part is strongly dependent on the topology of data center network.

3. Integration with BMS: The communication of the optimal set-points to the corresponding cooling systems will be governed by the data center BMS. Therefore, integrating the proposed SA with the BMS is critical. It involves working with multiple building automation protocols (e.g. BACnet).

4. Lack of Synergy: Data center is a complex cyber-physical system. Therefore, a cooling optimization tool like the proposed SA might destructively interfere with other data center components. For example, AdeptDC will increase air supply temperature to IT racks. That will increase server leakage power and server fan speed. Therefore, it is critical to examine whether the total energy consumption is reduced after the software deployment.

5. Humidity Control: As thermal management software appliance, the proposed workload-proportional cooling allocation technology would cause dynamic evolution of equipment inlet temperatures. While the proposed real-time analytics will ensure safe operating temperature, it important to control cooling air humidity through the BMS system. Failure to do so will cause imbalance in sensible heat vs. latent heat removal capacity and hinder effective cooling operation by the proposed technology.

For the phase I project, the primary focus will be on ensuring development of real-time analytics The secondary objectives will be BMS integration and getting access to IT device data. Once these operational pieces are developed, the focus will be shifted toward ensuring operational synergy for total energy usage reduction and humidity control.

Intellectual Property

The technology has been developed as a part of PhD research of Dr. Rajat Ghosh. Therefore, the technology is effectively a Georgia Tech property. A provisional patent (6563-PR-14) was filed on March 5, 2014. AdeptDC is planning to license the technology from Georgia Tech.

NSF Lineage

Dr. Rajat Ghosh (Entrepreneurial Lead), Dr. Yogendra Joshi (Principal Investigator), and Mark Davidson (Industry Mentor) took part in NSF Innovation Corps in the Arlington, VA cohort from July-September, 2014. The NSF award number is ICOL - 2506N02 (126050). The proposed innovation has been conceived as a part of Dr. Ghosh's doctoral research under the supervision of Prof. Joshi. This research was performed at the Consortium for Energy Efficient Thermal Management (CEETHERM) Data Center Laboratory, focused on data center energy efficiency. The activities of the lab have been supported by the following NSF awards:

Table 3: NSF awards supporting the progress of the project

NSF Award	Period	PI/co-PI	Program Manager
Site for I/UCRC on ES2	2013-2017	PI (Dr. Yogendra Joshi)	Drs. S.Priya, L. Hornak
Data Center Metrology Development	2010-2013	Co-PIs (Dr. Karsten Schwan, Dr. Yogendra Joshi)	

THE COMPANY/TEAM

AdeptDC co. is a software firm founded in October 2014 with the vision of providing an automated cooling optimization solution for data centers. The company was launched after participating in NSF I-Corps program in September 2014, which helped validate market need for the proposed SA. AdeptDC is a member of the Georgia Tech Advanced Technology Development Center (ATDC)—the nation's oldest, largest, and most successful university-affiliated incubator. As a member of ATDC, AdeptDC has access to a pool of successful serial entrepreneurs to serve as company mentors, connections to capital for future company growth, and educational programs to help the company prosper. Preliminary research was paid in part by a grant from the Georgia Research Alliance (GRA), which now enables AdeptDC to move forward with this SBIR Phase I study. Table 4 summarizes the key team members for AdeptDC co. These team members will be dedicated to work on this proposed Phase 1 SBIR project (July-December 2015). The key objectives for this phase will be to conduct a proof-of-concept study of the proposed technology.

Table 4: AdeptDC team members

Name	Designation	Responsibility
Rajat Ghosh	Chief Executive Officer	Principal Investigator
Yogendra Joshi	Chief Technical Mentor	Technical Advice
Karsten Schwan	Technical Advisor	Technical Advice
Mark Davidson	Chief Business Mentor	Strategy
Ada Gavrilovska	Vice President, Product Development	Computational Development
Dale Smith	Chief Engineer	Facility Implementation
Harold Brown	Vice President, Sales/ Marketing	Strategy and Market Development
Mary Miles	Chief Financial Officer	Financial Planning and Accounting
Zachary Allison	Lead Programmer	Software Development
Howard Hamilton	External Consultant	Real-time Control

Dr. Rajat Ghosh, CEO of AdeptDC co., will assume the responsibility of the Principal Investigator (PI) for the SBIR project. As a part of his doctoral research, Dr. Ghosh developed the real-time predictive analytics technology. He is developing a simulation-based prototype and directly interacting with customers to understand the market need. Dr. Ghosh has participated in NSF I-Corps program as the entrepreneurial lead and conducted 112 customer interviews. He is also working as the entrepreneurial lead for the GRA project that is helping AdeptDC to conduct early-stage market validation studies. He has published four journal papers and six conference papers in the related areas. His recent trade magazine article has been critically acclaimed in the press .

Prof. Yogendra Joshi, a renowned academic in electronics cooling, is a part of AdeptDC's executive board as a Chief Technical Mentor. He leads The CEETHERM Data Center Laboratory, which will be used for laboratory-scale testing. A pertinent subcontract has been already established with Georgia Tech in this regard. Prof. Joshi was the PI in the NSF I-Corps program for the related project and doctoral advisor for Dr. Ghosh. He was a co-founder of a Georgia Tech incubated company—Cool Clouds.

Prof. Karsten Schwan, Director of CERCS at Georgia Tech, is the Technical Advisor for the company. He is a renowned expert in high performance computing and real-time software tools.

Mr. Mark Davidson, Solution Executive for Cisco, is responsible for the business strategy development as a part of AdeptDC's executive board. He will bring nearly 30 years of experience in technology and sales for successful startups like JouleX. He was the industry mentor for the related NSF-Corps project.

Dr. Ada Gavrilvoska, Senior Research Faculty at Georgia Tech, is an expert in distributed computing, virtualization, and programmable network devices. If AdeptDC receives the award, she will be a part of the executive board as the Vice President of Product Development.

Mr. Dale Smith, Critical System Administrator at Internap, will work on integrating the SA to the BMS. He is a seasoned industry expert in data center controls. If AdeptDC receives the award, he will be a part of the executive board as the Chief Engineer.

Mr. Harold Brown, a successful serial entrepreneur and Mentor at Georgia Tech ATDC, will manage interactions with the customers, promotional strategies, investor relationships, and company brand management. He is a part of AdeptDC's executive board as the Vice President of Sales/ Marketing.

Mrs. Mary Miles, Mentor at Georgia Tech ATDC, will look after the financial affairs of the company as the Chief Financial Officer. She has considerable experience in working as a CFO for startup companies.

Mr. Zachary Allison, a successful entrepreneur and software developer, will be the lead software developer for AdeptDC. He has 15 years of experience, working as a Chief Technology Officer for startup companies. If AdeptDC receives the award, he will be a part of the executive board.

Dr. Howard Hamilton, a control system expert, will work as an external consultant and help AdeptDC in real-time analytics development.

Vision

AdeptDC's vision is that it will enable automated workload-proportional cooling in data centers. By doing so, it aspire to become one of the top players in the DCFM industry during the next five years. Given that AdeptDC will enter into a segmented and fairly risk-averse market, the adoption and the subsequent industry acceptance of the proposed SA will take some time. As a new company, AdeptDC does not have revenue. Table 5 shows estimated 5- year customer and revenue projections for AdeptDC.

Table 5: **Customer and revenue projections in first five years (assuming price remains same for 5 years)**

Year #	Customer #	Revenue Projection ($)
1	1	140K
2	3	420K
3	10	1,400K
4	25	3,500K
5	40	5,600K

TECHNICAL DISCUSSION AND R&D PLAN

The fundamental purpose of data center cooling is to remove waste heat from the facility. The cooling optimization problem pertains to minimizing data center cooling cost without increasing the risk of

service interruption and performance degradation. Any cooling mechanism that can keep the maximum IT device temperatures below a critical threshold within a reasonable operating budget can be a potential solution. One of the best solution strategies could be the workload-proportional cooling, which is defined as the cooling in sync with changes in IT workload. There are several technical challenges in implementing workload-proportional cooling. First, most data centers assess cooling need based on air temperature at: server inlets (15% of data centers), cooling hardware return sides (30%), cooling hardware supply sides (28%), and data room (27%) . As these locations are far away from the heat sources (e.g. CPU-s) and often strongly influenced by the convective airflow, the corresponding temperatures do not accurately anticipate the cooling demand. Second problem is related to inelastic cooling resource allocation mechanisms, in which the cooling set-points are kept static at the peak demand points. Given that the peak load happens only once or twice in a 24 hour period, a significant amount of cooling resources is wasted due to peak-provisioning. Third challenge in data center cooling management is lack of holistic optimization among several types of cooling equipment in data centers: room-level CRAC units, aisle-level in-row cooler units, rack-level rear door heat exchangers, and server fans. To reduce the overall cooling cost, it is important to ensure coordinated optimization across various cooling hardware. For example, the CRAC unit operating at a higher temperature set-point consumes less power; however, that would lead to higher server inlet temperatures and consequently trigger higher CPU leakage power and server fan power. Finally, absence of automated cooling set-point implementation means lack of synchronization between optimal set-point computation and their actual implementation. To solve the cooling optimization problem, the proposed SA will monitor in aggregate all data center device temperatures in near-real-time, compute cost-optimal cooling set-points with changing IT workloads in real-time, and automatically implement the optimal cooling set-points by communicating them to the corresponding cooling hardware.

* Figure 1 shows the architecture of the proposed SA that will solve data center cooling optimization problem. The central piece of the software appliance will be the real-time predictive analytics framework based on a machine learning based algorithm. The framework is fed by an IT device temperature data extraction system that communicates to multi-protocol (e.g. SNMP, IPMI)-based data center network. The IT device temperature data for different types of computational workloads and parametric cooling set-points will be extracted a priori and contained in a training database. Then, the training database will be filtered based on workload if the absolute deviation between prediction data and run-time IT device temperature data would fail to satisfy certain pre-assigned error criteria. The filtered data will be analyzed by a real-time algorithm to compute the optimal cooling set-point. The proprietary algorithm, which is based on a proper orthogonal decomposition (POD)-based iterative framework, is a logarithmic time algorithm . Therefore, it is particularly adept in efficiently handling large dataset in real-time. As shown inside the red rectangle (inset), the predictive analytics starts by determining IT device temperature data for the most cost-efficient cooling set-points, and then it assesses whether a reliability constraint is satisfied. The choice of the reliability constraint is dictated by data center Tier status and operating ranges of cooling equipment involved. If the reliability constraint is not satisfied, then an iterative procedure that transfers the computation to the next cost-effective cooling set-points is invoked. The robustness of the algorithm was validated in . The predicted temperature data will be compared with the runtime temperature data. If the maximum absolute deviation is within a pre-assigned error limit, the optimal cooling set-points will be transferred to the cooling devices by the

installed cooling network drivers (e.g. BACnet driver) and management platforms (e.g. BMS) to automatically implement optimal cooling. Otherwise, a subset of the training data set corresponding to another workload type will be chosen. With the data-driven algorithm, the framework is flexible enough to accommodate various types of IT devices, cooling equipment, and communications protocols (e.g. BACnet, Modbus).*

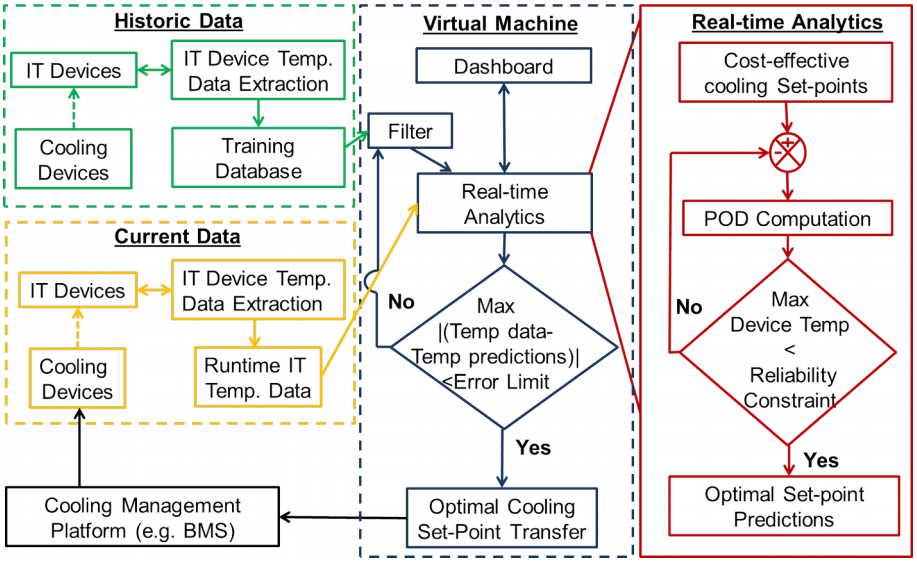

Figure 1: The architecture of the proposed cooling optimization software appliance

* The effectiveness of the proposed real-time predictive analytics was validated in laboratory-scale data center (600 sq. ft., 150 kW heat load, 14,000 CFM rack flow rate, 7x2 alternating cold aisle/hot aisle architecture, 6000 CFM tile flow rate, under-floor plenum supply and overhead plenum return) with four simulated computing workloads, namely cloud, batch, enterprise, and HPC. Commercially available vSphere software was used for the workload simulation purpose. Using commercially available PI system, the CPU temperatures are compiled for these four simulated workloads and 12 different cooling environments: defined by the set-points for a 150 kW CRAC unit for cooling air delivery into the plenum at four specified temperatures (17, 21, 25, 29 ^{0}C) and 18 kW rack integrated air to chilled water rear door heat exchanger (RDHx) unit at three pressure set-points for supplied chilled water (4, 7, 10 psi). The measurement time window (T) was 3000 s and sampling period was 75 s. The reliability constraint in this case study is defined as the CPU temperature is less than or equal to 65 ^{0}C. Figure 2 shows a representative cloud workload which consists of two components: a periodic component (representative of typical diurnal variations) and a spike-like increment (representative of flash crowd). Figure 2 also shows the recommendations (green lines) for CRAC supply temperature and RDHx pressure set-points based on the proposed analytics framework. As the CRAC supply temperature goes down and the RDHx operating pressure moves up with the increase in CPU utilization and vice versa, the workload-proportional nature of the proposed real-time predictive analytics framework is demonstrated—higher CPU utilization means more waste heat and higher cooling demand. The recommended set-points are compared with the industry standard set-points (18 ^{0}C and 12 psi) and cost-saving benefits are assessed [13]. The average cost savings for CRAC is calculated to be equal to 19% and that for RDHx is 50%. Similar studies have been conducted for other types of workloads . The cost-savings numbers for other workloads are shown in Table 2. A close scrutiny of Figure 2 also indicates coordination between two types of cooling units (i.e. CRAC and RDHx) and holistic cooling energy optimization. Figure 2

indicates that CRAC set-points do not change during 0.8T-0.84T (flash crowd); however, RDHx pressure set-point rapidly jumps to match the high cooling demand. It means the proposed control mechanism will first utilize low-power RDHx fully before lowering high-power CRAC unit. *

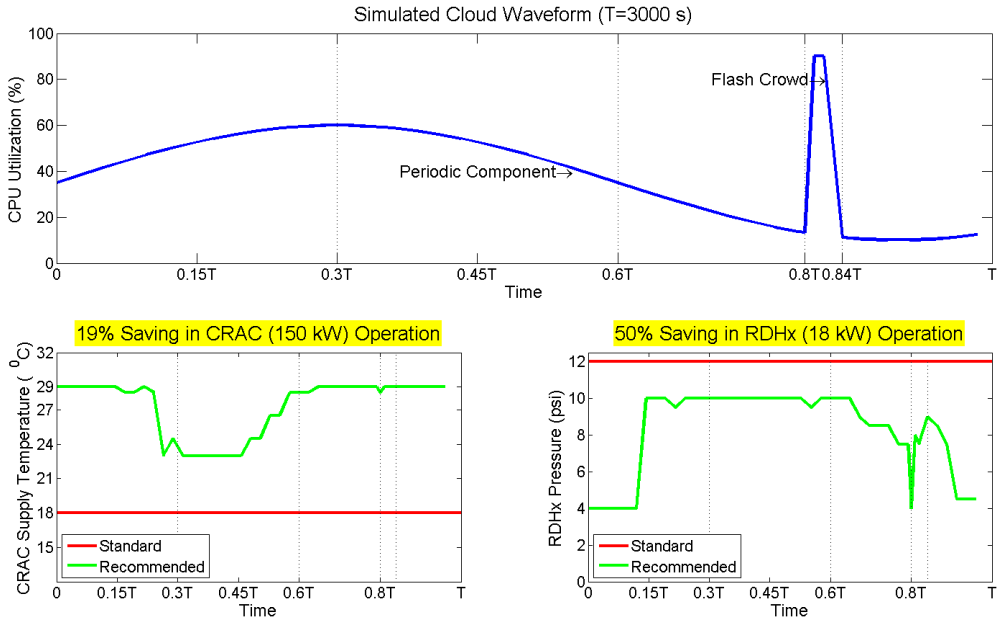

Figure 2: Workload-proportional cooling set-point computation for a cloud workload

R & D Objectives

The goal of the proposed SBIR Phase-I research is to conduct a feasibility study, which will test the potential of the proposed technology in a small scale. Therefore, the R&D objectives of this project are:

1. Minimal Viable SA Development: As identified in Figure 1, there are six key technical components for AdeptDC: IT data extraction system, training database, filter, real-time predictive analytics, dashboard, and cooling automation system. The key focus will be to develop these capabilities using various open-source tools.

2. Component Testing**:** Six developed components will be tested. Key assessment criteria will be application runtime, execution stability, and operational fidelity. Suitable modifications will be made until the performance expectations are met.

3. Integration**:** Different components of the software system will be aggregated into the SA (similar to Figure 1). The assessment criteria will be end-to-end data flow, computational time, and prediction accuracy. At the end of this step, the SA will be ready for the deployment.

4. System-level Testing**:** The SA will be deployed in a laboratory-scale data center such the CEETHERM Lab at Georgia Tech to evaluate its performance. It will be tested in various cooling and computational environments to assess its operational characteristics. [Key Questions: Is the appliance behaving reliably with a reasonable efficiency and accuracy? Are there any deployment risks? How to make the system more efficient, reliable, and user-friendly?]

5. Performance Benchmarking: The refined SA will be deployed in a real data center site such as the Georgia Tech OIT data center to assess its cost-saving potential, and understand real-life deployment issues. The software features will be revised to rectify apparent shortcomings. [Key Questions: Does the software perform reliably? What are the cost savings figures? Does the software ensure coordinated optimization across multiple cooling units?]

Technical Milestones

Table 6 shows the technical milestones to reach the identified R&D objectives:

Table 6: **The technical milestones for reaching R&D objectives**

R&D Objectives	Technical Milestones
1.Minimal Viable SA Development	(a) IT Device Temperature Data Extraction System (b) Training Database (c) Filter (d) Real-time Analytics (e) Dashboard (f) Cooling Automation System
2. Component Testing	(a) Testing (b) Test Result Database Compilation and Adjustment
3.Integration	(a) Developing Bridging Protocols
4.System-level Testing	(a) Laboratory-scale Deployment (b) Exhaustive Testing (c) Test Result Compilation and Refinement
5.Performance Benchmarking	(a) Live Deployment (b) Continuous Performance Monitoring (c) Evaluation and Refinement

R&D Plan

This section pertains to plan of actions to reach the proposed R&D objectives:

1. *Minimal Viable SA Development*

(a) IT Device Temperature Data Extraction System: This system will communicate to the IT devices in the data center network and pull out the device temperature data (e.g. CPU temperatures). It is planned that OpenStack-based software platform will be used to interact with multi-protocol-based (e.g. SNMP, IPMI etc.) network management systems in data centers. Its effectiveness will be tested on the basis of how efficiently (in real-time) it can poll device temperature data.

(b) Training Database: This piece will store the training database, comprised of IT device temperatures, for different IT workloads and parametric cooling set-points. It will be developed in MySQL.

(c) Filter: This component will filter the training database based on workloads if the maximum absolute deviation between runtime IT data and corresponding predictions would fail to satisfy a pre-assigned error limit. This capability will be developed in Python.

(d) Predictive Analytics: This will take filtered training data and process in real-time to recommend the optimal cooling set-points. A real-time predictive analytics framework is being developed under the

ongoing GRA project. The proposed SBIR Phase-I project will work on improving its computational efficiency, prediction accuracy, and data flow. This capability will be developed in Python. Its performance will be assessed based on computational efficiency and prediction accuracy.

(e) Dashboard: This will be an input/ output (I/O) port where users can plug in their efficiency and accuracy requirements. A web-based dashboard is being developed under the ongoing GRA project. The proposed SBIR Phase-I project will design an enhanced user interface, which will be developed in JavaScript, CSS, and PHP. It will display the real-time evolution of IT device temperatures and cooling set-points. The dashboard will include user-interactive switching capability with which users can easily shift between manual and automated control modes. Its effectiveness will be assessed by its ease of use and intuitiveness.

(f) Cooling Automation System: This capability will transfer the optimal cooling set-points to the cooling management platform (e.g. BMS-s and PLC-s) which will in turn actuate optimal cooling. This transfer will require encoding optimal cooling set-points into the building control driver (e.g. BACnet driver). The accuracy of the transferred signal and speed of transmission will be assessed.

2. *Component Testing*

(a) Testing: All six components will be tested under different operating conditions. The operating ranges and pertinent operating parameters of these components will be assessed. Several operating points will be chosen for exhaustive component testing. The test data, including computational runtime, stability, and fidelity, will be noted in a database.

(b) Evaluation and Adjustment: The stored data will be analyzed statistically to understand the component characteristics. Necessary adjustments will be made to enhance the component functionalities and reliability.

3. *Integration*

(a) Developing Bridging Protocols: This step will be dedicated for coupling six different functional components. It will be largely accomplished in Linux platform and various open source tools will be used for the integration. The feasibility of this step will be determined by seamless operation of the resulting integrated system.

4. *System-level Testing*

(a) Laboratory-scale Deployment: The integrated solution will be deployed in a laboratory-scale data center environment at the CEETHERM Data Center Laboratory at Georgia Tech. The input side will involve connecting the developed IT data extraction system with the existing SNMP network managers. This system will fetch training database and runtime data, comprised of IT device temperatures. The output will entail communicating the cooling set-point outputs to the cooling management (e.g. BMS) units via building automation communication protocols such as BACnet. That will essentially entail programming building automation drivers (e.g. BACnet drivers). The feasibility of this step is governed by ease of operational connectivity and compilation of training data.

(b) Exhaustive Testing: The software appliance will be tested under various operating conditions governed by different underlying parameters such as simulated workload shapes, cooling equipment choices, and user-defined efficiency and fidelity criteria. The workload shapes can be regulated by

commercially available vSphere software. Cooling equipment can be selectively connected or disconnected. The user inputs can be set for various choices of efficiency and fidelity criteria.

(c) Test Result Compilation and Refinement: A full-factorial test will be conducted to cover an exhaustive parametric space. The efficiency and reliability of the system will be noted. Systemic statistical analyses will be performed to understand the system characteristics. The issues revealed (e.g. performance consistency) during the system-level testing will be fixed in this step.

5. *Performance Benchmarking*

(a) Live Deployment: This step will involve deploying the SA in a live data center environment such Georgia Tech OIT Data Center, which serves Georgia Tech and few other colleges in Atlanta. The deployment will entail connecting the IT extraction system to the data center network managers. It will also require connecting output port to the BMS system. The deployment activity will require working around the data center network security system. Ease of use and rapid deployment will be the key feasibility criteria for this step. This step will also involve a priori compilation of the training database for the real-time analytics. The feasibility of this step will be judged by the ease of IT data access, training data compilation, and transferring optimal cooling data to the control hardware.

(b) Continuous Performance Monitoring: The SA will be deployed in the production facility for an extended period of time (e.g. a week), so that it could be exposed to the typical duty-cycle of the data center. While benchmarking variables will be stored in a database and monitored graphically, the maximum CPU temperature of the facility will be noted. By design, the maximum device temperature should always be below a critical limit. Otherwise, an alert will be created.

(c) Evaluation and Refinement: In this step, statistical analyses will be conducted on monitored data to understand the performance characteristics of the SA in a real data center environment. A key emphasis of this step will be to scrutinize the differences between SA operations in lab and in a real facility. The SA will be refined to fix the potential issues such as CPU overheating.

Proof of Concept Criteria for the Developed Technical Solution

The small-scale prototype will prove the proposed concepts if the following criteria are met:

- Cooling energy saving by at least 5% for HPC workload and 20% for non-HPC workload
- Cooling capacity enhancement by at least 20%
- Co-ordinated optimization between different cooling units
- The maximum IT device temperature always under the reliability constraint

Based on these criteria, if above technical objectives and milestones are met, AdeptDC will have achieved technical feasibility and will be ready to begin Phase-II research.